Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation

Russell, Joanne; Mascher, Martin; Dawson, Ian K.; Kyriakidis, Stylianos; Calixto, Cristiane; Freund, Fabian

Published in:
Nature Genetics

DOI:
10.1038/ng.3612

Publication date:
2016

Document Version
Peer reviewed version

Supplementary Note

Read mapping, read depth analysis and SNP calling

We created a custom reference sequence in order to be able to use GATK with the highly fragmented WGS assembly of cv. Morex. The contigs of the Morex assembly were concatenated into 25 pseudoscaffolds. Adjacent contigs were separated by 500 'Ns'. A FASTA file with the pseudoscaffolds and a BED file with positions of the original WGS contigs is available from DOI. Mapping files were converted to BAM format, sorted and subjected to duplicate removal with PicardTools version 1.100[a]. Read depth statistics were calculated with GATK DepthOfCoverage. Realignment around indels, base score recalibration and variant calling were performed with GATK version 2.7.4 [b]. As GATK base quality score recalibration (BQSR) requires a comprehensive genome-wide set of variants, two rounds of preliminary SNP calling were performed where high-confidence scores of prior rounds were used as input for BQSR. In the first round of variant calling, multiple sample calling across all sequenced samples was performed with the SAMtools/bcftools pipeline[1]. The command "SAMtools mpileup" was used with the parameter "-D" to record per-sample read depth. Otherwise, default parameters were applied. Prior to the second round of variant calling, BQSR was run with high-confidence SNPs determined in the first round. Then, SNPs were called with the GATK UnifiedGenotyper. Variant filtering was performed as described below for the final variant set. Prior to the third and final round of variant calling with the GATK UnifiedGenotyper, BQSR was performed using the intersection of SAMtools (first round) and GATK (second round) as well as variant positions detected by sequencing two bi-parental mapping populations[2]. The final variant calls were filtered with an AWK script available as Supplementary Text S3 of Mascher *et al* [2]. Genotype calls were considered successful if read depth and the genotype quality score were both ≥ 10. Genotype calls not passing these filters were set to missing. SNP positions with more than 90% heterozygous calls or more than 20% missing genotype calls were discarded. Only bi-allelic variants (SNPs/indels) were considered. The resultant SNP matrix was imported into the R statistical environment[c] and further filters were applied. Variants that were fixed differences between *H. vulgare* and *H. bulbosum* (i.e. monomorphic in *H. vulgare*) were discarded. Next, both alleles of a variant were required to occur in at least one individual in the homozygous state. Finally, SNPs with less than 95% present genotype calls were discarded.

Annotation of false positive SNP sets

The raw Illumina reads were quality trimmed to a base quality of 20 from both ends with Trimmomatic version 0.30[3]. Only correctly paired reads longer than 50 bp were used for further processing. Reads were then mapped to the reference sequence with Bowtie2[4], using the "--very-sensitive" flag to improve mapping accuracy. An equivalent of only two mismatches per read was allowed. The mismatch rate in Bowtie2 is controlled with the "--score-min" parameter, which is computed for each read separately as a function of the read's length (see the Bowtie2 manual[d]). The score-min parameter used here was "L,0,-

0.12". The first parameter ("L") specifies a linear relationship between read length and the number of mismatches. The second parameter (0) is the *y* intercept, and the third parameter represents the coefficient for the slope of the regression equation used. To calculate this coefficient for our dataset, we divided the intended maximum mismatch score (obtained by multiplying the maximum number of mismatches in a full length, untrimmed read by the default penalty -6) by the read length, giving a coefficient of -0.12 ((2 *-6)\100). The resulting BAM files were sorted and merged using SAMtools version 0.1.18[1] to produce a single BAM file for variant calling.

Variants were called on the combined BAM file with Freebayes version 0.9.9[5] using the following set of parameters:

--min-alternate-count 3 --min-alternate-total 3 --min-alternate-fraction 0.9 --no-mnps --min-mapping-quality 20 --min-base-quality 20 -v contig_1.vcf --no-population-priors --ploidy 2

Although FreeBayes includes some scripts that allow it to run across multiple CPUs or cores of a single machine, there is no way of running it across a cluster of machines. To achieve this, we implemented a Java application to split jobs into smaller chunks that could then run independently on separate cluster nodes. The idxstats routine of SAMtools[1] was first used to determine both the number of contigs in the BAM file and how many of them had a read count greater than one. A chunking algorithm was then used to split the contigs into batches, with each batch being assigned its own CPU for processing. Using batches of contigs assisted in lowering job scheduling overheads, as it is too inefficient to simply assign one contig per CPU when there are several million to process. The final step was to concatenate the VCF output files from each batch into a single, ordered result. This allowed us to fully utilize our high-performance compute resource, and significantly reduced the runtime of the job. The entire dataset was processed in under 24 hours, down from an estimate of three months without parallelization. This initial call set was then filtered with custom Java code to remove variants with a quality score of less than 20, and also those variants where more than 20% of samples were heterozygous.

We then annotated the SNPs with respect to Illumina systematic sequencing error. This type of error is associated with certain sequence motifs and can occur anywhere in a read[6,7]. The resulting errors are typically associated with extreme strand bias, i.e. the alternate allele occurs exclusively or almost always on either forward or reverse reads only. Another reliable indicator is the base quality of the alternate alleles, which is consistently lower than that of the reference allele, but usually still high enough for a variant not to be filtered out on grounds of poor base quality. Based on this characteristic we developed custom Java code which calculates for each variant the percentage difference between the mean alternate base quality and the mean reference base quality. The cut-off used here was 10%. A total of 856,008 SNPs were annotated as being due to Illumina systematic sequencing error, and these were subtracted from the final SNP set described in the main manuscript.

We also annotated a set of variants as reference sequence assembly errors. We mapped reads from the Morex exome capture sample to the Morex reference sequence and then called variants using Freebayes with the following parameters:

--min-alternate-count 3 --min-alternate-total 3 --min-alternate-fraction 0.9 --no-mnps --min-mapping-quality 20 --min-base-quality 20 --no-population-priors --use-reference-allele --ploidy 2

This resulted in a call set containing 101,025 SNPs, all of which were subtracted from the final SNP set described in the main manuscript. The variant call sets were spot-checked for consistency and correctness before and after the false positive removal using the Tablet assembly viewer[8].

SNP validation

We compared variant calls between technical replicates and checked concordance with array-based genotyping and whole-genome sequencing datasets. Exome capture and sequencing were performed for thirteen samples starting from the same DNA. The average concordance of genotype calls at SNP positions between replicated samples was 99.5%. If heterozygous calls in either replicate were excluded, the concordance increased to 99.9%.

Of the accessions sequenced in the present study, 148 had been genotyped at 7,854 SNP positions using an Illumina 9k iSelect SNP chip[9]. A total of 3,882 SNPs on the chip were located in target regions of the exome capture assay. Of these, 3,548 (91%) were also called as variant positions from the exome sequencing data. The agreement of genotype calls made from array data and from exome sequencing data was 98.4%. If heterozygous calls in either the array or the sequence data were excluded, the concordance increased to 99.1%.

Next, we compared genotype calls between exome capture and whole genome datasets. High-coverage whole genome shotgun (WGS) data had been previously obtained for two cultivars (Bowman: 35-fold coverage, Barke: 30-fold coverage). SNPs had been called using a pipeline consisting of BWA and SAMtools. Two exome capture libraries were sequenced either alone (high coverage) or in eight-fold multiplex (low-coverage) on one HiSeq2000 lane and included in the present study. The proportion of variants identified between Bowman and Morex from exome capture data also present in the WGS data was 96.7% for the low-coverage dataset and 96.1% for the high-coverage dataset. The respective numbers for Barke were 91.9% and 90.5%.

Finally, we compared our variant calls for Barke and Morex to sequencing data of individuals of a bi-parental mapping population derived from a cross between these two genotypes (POPSEQ). SNPs detected by POPSEQ are true variants with very high confidence as these SNPs could be mapped genetically in the recombinant progeny. A total of 103,101 genetically-mapped POPSEQ SNPs were located in target regions of the exome capture data (i.e. regions having 10-fold read coverage in at least 95% of all accessions). Of these SNPs,

84,283 (81.7%) were present in the final exome capture call sets; 95,471 (92.6%) had been present in the unprocessed GATK output. Factors other than raw sequence coverage, such as mapping quality, base quality, strand bias or proximity to indels or regions with low coverage at the border of capture targets, may have resulted in not calling or filtering out high-confidence POPSEQ SNPs.

Prediction of functional effects

To avoid annotation artifacts due to fragmented gene models, we predicted functional effects only for SNPs located in genes whose protein sequence is nearly completely represented in the Morex WGS assembly. Protein sequences were aligned to the genomic contigs with the program exonerate[10]. Genes were considered near-complete if 98% of their protein sequences could be aligned to the genomic sequence. A total of 18,039 (74%) out of 24,243 high-confidence genes positioned on the Morex WGS assembly had near-complete ORFs. Functional effects were predicted with SnpEff version 3.4[11], using a total of 705,471 SNPs located in exons of near-complete genes as input. A summary of the assignment of SNPs to annotated gene models is given in **Supplementary Table 2**.

References

1. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
2. Mascher, M. *et al.* Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J. 76, 718-727 (2013).
3. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**, 2114-2120 (2014).
4. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Meth*. **9**, 357-359 (2012).
5. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 http://arxiv.org/abs/1207.3907 (2012).
6. Nakamura, K. *et al*. Sequence-specific error profile of Illumina sequencers. *Nucl. Acids Res.* **39**, e90 (2011).
7. Meacham, F. *et al*. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).
8. Milne, I *et al*. Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinformat.* **14**, 193-202 (2013).
9. Comadran, J. *et al.* Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. Nat. Genet. 44, 1388-1392 (2012).
10. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

11. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* (Austin) **6**, 80-92 (2012).

URLs

a. http://picard.sourceforge. Net
b. http://www.broadinstitute.org/gatk/guide/best-practices
c. http://www.r-project.org
d. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#bowtie2-options-score-min

Supplementary Tables

Supplementary Table 1. Information on 267 exome-captured barley accessions (SEE SEPARATE FILE)

Supplementary Table 2. Assignment of SNPs to annotated gene models

Supplementary Table 3. Individual and synthetic environmental variables used for environmental association analyses

Supplementary Table 4. Homologues of *Arabidopsis* flowering-associated genes

Supplementary Table 5. SNPs present in flowering-associated genes (SEE SEPARATE FILE)

Supplementary Table 6. SNPs equally or more associated with days to heading than *HvCEN* SNP 274284:918, across two common garden trials

Supplementary Table 2. Assignment of SNPs to annotated gene models**.**

HC exon	745,815 (44.2 %)
HC intron	204,105 (12.1 %)
LC exon	342,690 (20.3 %)
LC intron	36,032 (2.1 %)
HC/LC upstream/downstream ± 1 kb	108,227 (6.4 %)
intergenic	251,938 (14.9 %)

Footnote: A full description of the composition of the exome capture array, from where the above figures are derived, is given in Mascher *et al*[1]. In summary, the capture contains 61.6 Mbp of coding sequence target from the genome assembly of the cultivar Morex, publicly available full-length cDNAs and *de novo* assembled RNA-Seq consensus sequence contigs. Intron, intergenic, 5' and 3' sequences are therefore generally sampled as a result of their adjacency to exon targets.

Supplementary Table 3. Environmental variables for association analyses.

Code	Description	Cluster (individual or synthetic)
BIO1	Annual mean temperature	BIO1+BIO6+BIO11
BIO2	Mean diurnal range (mean monthly [max t. - min t.])	Individual
BIO3	Isothermality (BIO2/BIO7) (* 100)	Individual
BIO4	Temperature seasonality (SD *100)	BIO4+BIO7
BIO5	Maximum temperature of warmest month	PET+BIO5+BIO9+BIO10
BIO6	Minimum temperature of coldest month	BIO1+BIO6+BIO11
BIO7	Temperature annual range (BIO5-BIO6)	BIO4+BIO
BIO8	Mean temperature of wettest quarter	Individual
BIO9	Mean temperature of driest quarter	PET+BIO5+BIO9+BIO10
BIO10	Mean temperature of warmest quarter	PET+BIO5+BIO9+BIO10
BIO11	Mean temperature of coldest quarter	BIO1+BIO6+BIO11
BIO12	Annual precipitation	BIO12+BIO13+BIO16+BIO1
BIO13	Precipitation of wettest month	BIO12+BIO13+BIO16+BIO1
BIO14	Precipitation of driest month	GAI+BIO14+BIO17
BIO15	Precipitation seasonality (coef. variation)	Individual
BIO16	Precipitation of wettest quarter	BIO12+BIO13+BIO16+BIO1
BIO17	Precipitation of driest quarter	GAI+BIO14+BIO17
BIO18	Precipitation of warmest quarter	BIO12+BIO13+BIO16+BIO18
BIO19	Precipitation of coldest quarter	Individual
Solar1	Solar radiation January	Latitude+Solar1-4+Solar8-12
Solar2	Solar radiation February	Latitude+Solar1-4+Solar8-12
Solar3	Solar radiation March	Latitude+Solar1-4+Solar8-12
Solar4	Solar radiation April	Latitude+Solar1-4+Solar8-12
Solar5	Solar radiation May	Solar5-
Solar6	Solar radiation June	Solar5-
Solar7	Solar radiation July	Solar5-7
Solar8	Solar radiation August	Latitude+Solar1-4+Solar8-12
Solar9	Solar radiation September	Latitude+Solar1-4+Solar8-12
Solar10	Solar radiation October	Latitude+Solar1-4+Solar8-12
Solar11	Solar radiation November	Latitude+Solar1-4+Solar8-12
Solar12	Solar radiation December	Latitude+Solar1-4+Solar8-12
GAI	Global aridity index	GAI+BIO14+BIO17

PET	Annual global potential evapotranspiration	PET+BIO5+BIO9+BIO10
Long	Longitude	Individual
Lat	Latitude	Latitude+Solar1-4+Solar8-12; individual
Altitude	Elevation	Individual

Supplementary Table 4. Functional homologues of *Arabidopsis* flowering-associated genes.

Genes	*Arabidopsis thaliana*	*Hordeum vulgare*
LHY/CCA1	At1g01060 (*LHY*) At2g46830 (*CCA1*)	Hvcontig_1567295 (*HvLHY)*
LUX	At3g46640 (*LUX*)	Hvcontig_2548416 (*HvLUX*)
ELF3	At2g25930	Hvcontig_80895/67536 (*HvELF3*)
GI	At1g22770	Hvcontig_58270/1580005 (*HvGI*)
TOC1	At5g61380	Hvcontig_37494 (*HvTOC1*)
PRR5(9)/PRR9(5)	At5g24470 (*PRR5*) At2g46790 (*PRR9*)	Hvcontig_46739 (*HvPRR59*) Hvcontig_41351 (*HvPRR95*)
PRR7/37	At5g02810 (*PRR7*)	Hvcontig_94710 (*HvPPD-H1/PRR37*)
ZTL	At5g57360 (*ZTL*)	Hvcontig_273830 (*HvZTLa*) Hvcontig_158755 (*HvZTLb*)
FKF1	At1g68050	Hvcontig_38586 (*HvFKF1*)
GRP7	At2g21660	Hvcontig_1578172 (*HvGRP7a*) Hvcontig_43832/46175 (*HvGRP7b)*
CO	At5g15840 (*CO*) At5g15850 (*COL1*) At3g02380 (*COL2*)	Hvcontig_138334 (*HvCO1*) Hvcontig_6805 (*HvCO2*)
FT	At1g65480	Hvcontig_54983 (*HvFT1* or *VRN-H3*) Hvcontig_1558556/136243 (*HvFT2*)
ELF4-like3	At2g06255 (*ELF4-like3*)	Hvcontig_42805 (Hv*ELF4-like*[*A*])
CEN	AT2G27550	Hvcontig_274284 (*HvCEN*)

Supplementary Table 6. SNPs equally or more associated with days to heading than *HvCEN* SNP 274284:918, across two common garden trials.

SNP
135011:2287
137614:5689
1562837:4807
1562837:4905
2548323:3116
2548323:4037
2548323:5097
39745:3186
39745:3417
40046:7262
41141:5038
42711:4614
42711:4718
42711:4966
44918:10916
44918:2629
44918:4418
45846:1433
50062:3893
50062:4125
53050:5340
55785:3010
55785:3066
5702:4724
5702:4883
59659:4243
59659:4259

Supplementary Tables references

1. Mascher, M. *et al*. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J*. **76**, 494-505 (2013).
2. Turner, A., Beales, J., Faure, S., Dunford, R.P. & Laurie, D.A. The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Science* **310**, 1031-1034 (2005).
3. Jones, H. *et al.* Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent. *Mol. Biol. Evol.* **25**, 2211–2219 (2008).

Supplementary Figures 1 - 31

Supplementary Fig. 1. Minor allele frequency (MAF) distributions of SNPs in barley groups.

Supplementary Fig. 2. Linkage disequilibrium (r^2) in wild and domesticated barleys**.**

Supplementary Fig. 3. Principal component analyses of geo-referenced barley group accessions.

Supplementary Fig. 4. Spatial autocorrelation analyses of geo-referenced spontaneum and landrace accessions.

Supplementary Fig. 5. Differentiation (F_{ST}) between two-row and six-row phenotypes within the geo-referenced landrace barley group accessions.

Supplementary Fig. 6. Test for selective sweeps across geo-referenced spontaneum and landrace barley group accessions.

Supplementary Fig. 7. X^TX and Bayes factor results.

Supplementary Fig. 8. Regression plots of individual components of the synthetic variable (PET+BIO5+BIO9+BIO10) with PC1 days to heading and height.

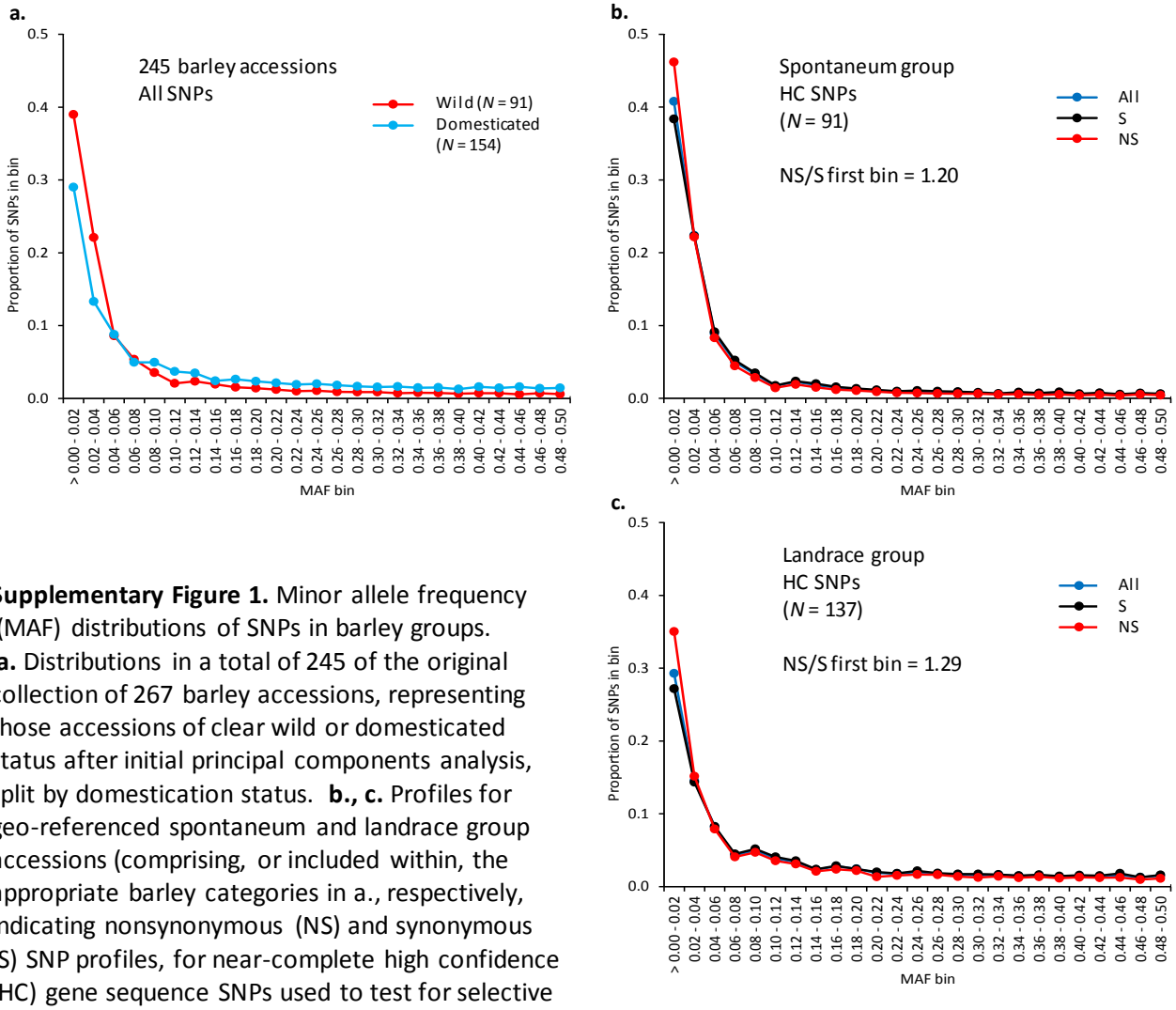
Supplementary Fig. 9. Proposed schematic gene network of barley flowering-associated genes superimposed on the *Arabidopsis* model.

Supplementary Fig. 10. Distribution of flowering-associated genes used in our study across chromosomes on the barley genome.

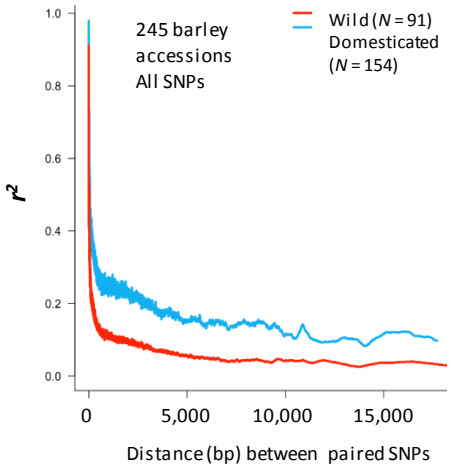
Supplementary Fig. 11. Spatial autocorrelation analyses of genic SNPs for flowering-associated genes.

Supplementary Fig. 12 - 30. Information on haplotype profiles for 19 flowering-associated genes in spontaneum and landrace groups.

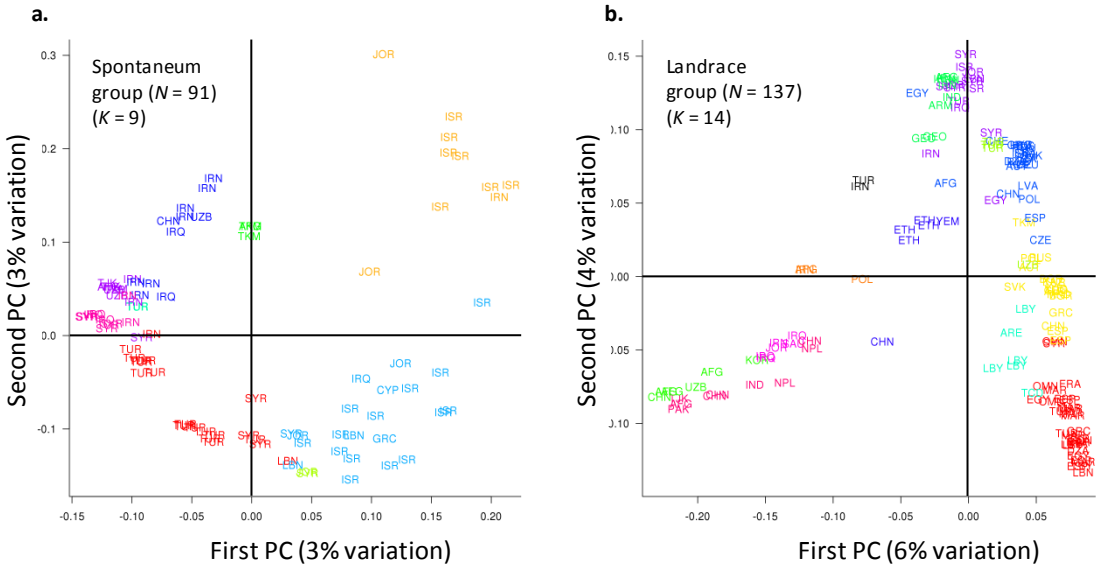
Supplementary Fig. 31. Fractional sNMF assignments for 228 geo-referenced accessions based on all genic SNPs pooled across 19 flowering-associated genes.



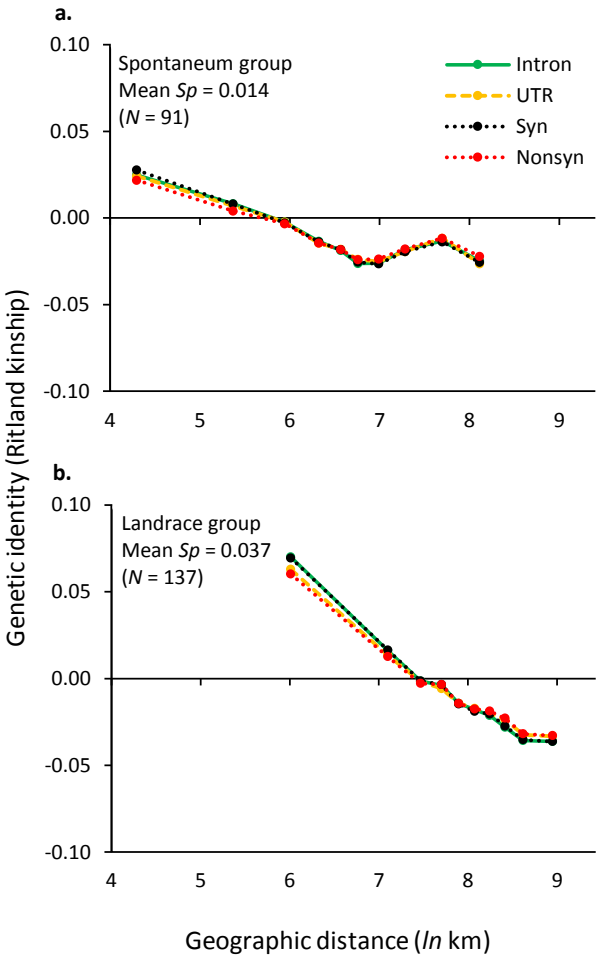
Supplementary Figure 1. Minor allele frequency (MAF) distributions of SNPs in barley groups. **a.** Distributions in a total of 245 of the original collection of 267 barley accessions, representing those accessions of clear wild or domesticated status after initial principal components analysis, split by domestication status. **b., c.** Profiles for geo-referenced spontaneum and landrace group accessions (comprising, or included within, the appropriate barley categories in a., respectively, indicating nonsynonymous (NS) and synonymous (S) SNP profiles, for near-complete high confidence (HC) gene sequence SNPs used to test for selective sweeps and environmental associations. The proportions of SNPs in consecutive bins with a frequency increment of 0.02 are given. Profiles reveal an excess of low MAF SNPs in wild/spontaneum barley compared to domesticated/landrace barley, as well as an excess of low MAF NS SNPs compared to S SNPs in spontaneum and landrace groups. The excess of NS SNPs, calculated as the NS/S ratio of proportions for the first MAF bin, representing singletons for the spontaneum group and mostly singletons for the landrace group, is higher in the latter case.



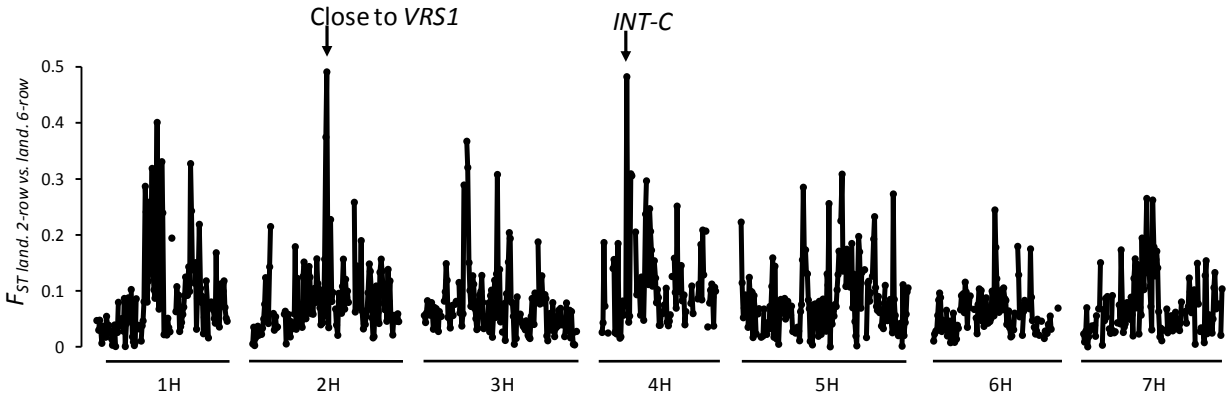
Supplementary Figure 2. Linkage disequilibrium (r^2) in wild and domesticated barleys. Estimates are shown for a total of 245 of the original collection of 267 barley accessions, representing those accessions of clear wild or domesticated status after initial principal components analysis, split by domestication status. Values were determined for all pairs of SNPs on contigs and plotted as a function of distance between SNP pairs. Rolling medians of 5,000 SNPs were generated to produce a smoothed decay curve. Estimates decay more rapidly and reach a lower basal level for wild barley.



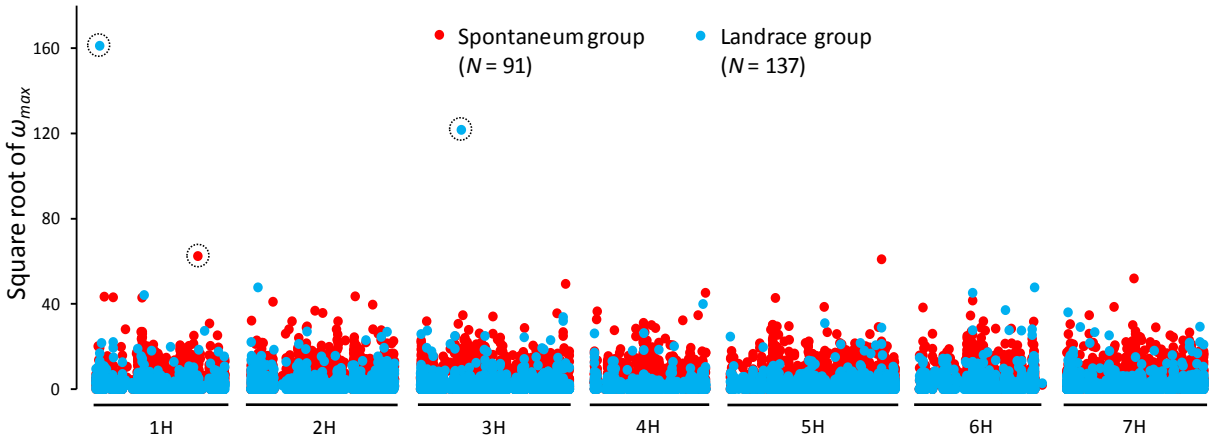
Supplementary Figure 3. Principal components analyses of geo-referenced barley group accessions. **a.** Spontaneum. **b**. Landrace. In each case, the first two principal axes are given. *K*-group assignments (with no individual admixture, see **Online Methods**) are indicated by color coding. Country of origin is given by three-letter country codes according to FAO designations.



Supplementary Figure 4. Spatial autocorrelation analyses of geo-referenced **a.** spontaneum and **b.** landrace accessions. Profiles are based on representative random systematic subsampling of 5,000 genome-wide SNPs for each of four SNP categories. Patterns for SNP categories within barley groups correspond closely, while the overall fall-off in identity with distance is highly significant ($P < 0.001$) for both barley groups with all four SNP categories. Values of the *Sp* statistic (mean across SNP categories for barley groups), which quantifies spatial genetic structure based on the regression slope of a profile and the kinship coefficient of the first distance class, indicate a more rapid fall-off in genetic identity with geographic distance in the landrace group.

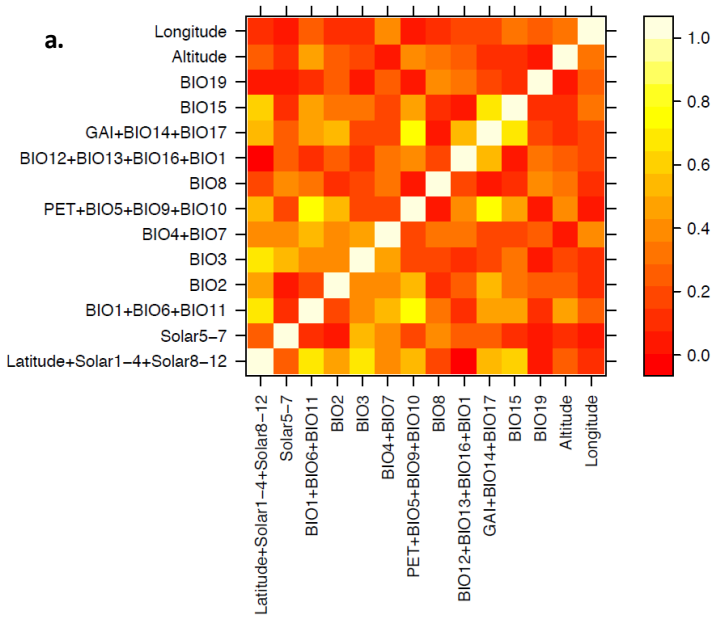


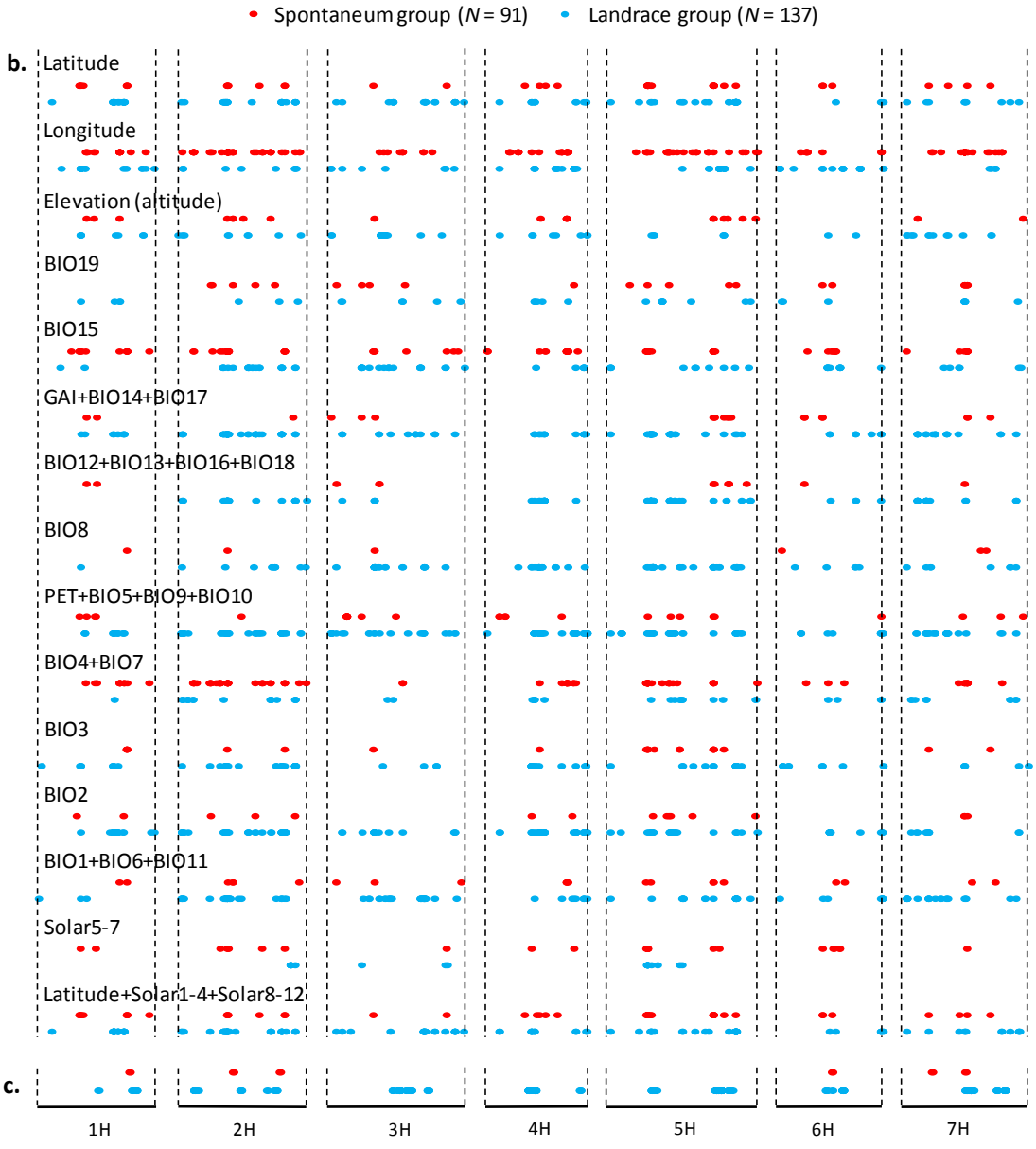
Supplementary Figure 5. Differentiation (F_{ST}) between two-row (N = 55) and six-row (N = 82) phenotypes within the geo-referenced landrace barley group accessions. Values are for 1 cM bins but otherwise calculated with the same approach as differentiation between spontaneum and landrace barley accession groups (**Fig. 2c**). Features noted in the text are indicated by arrows.

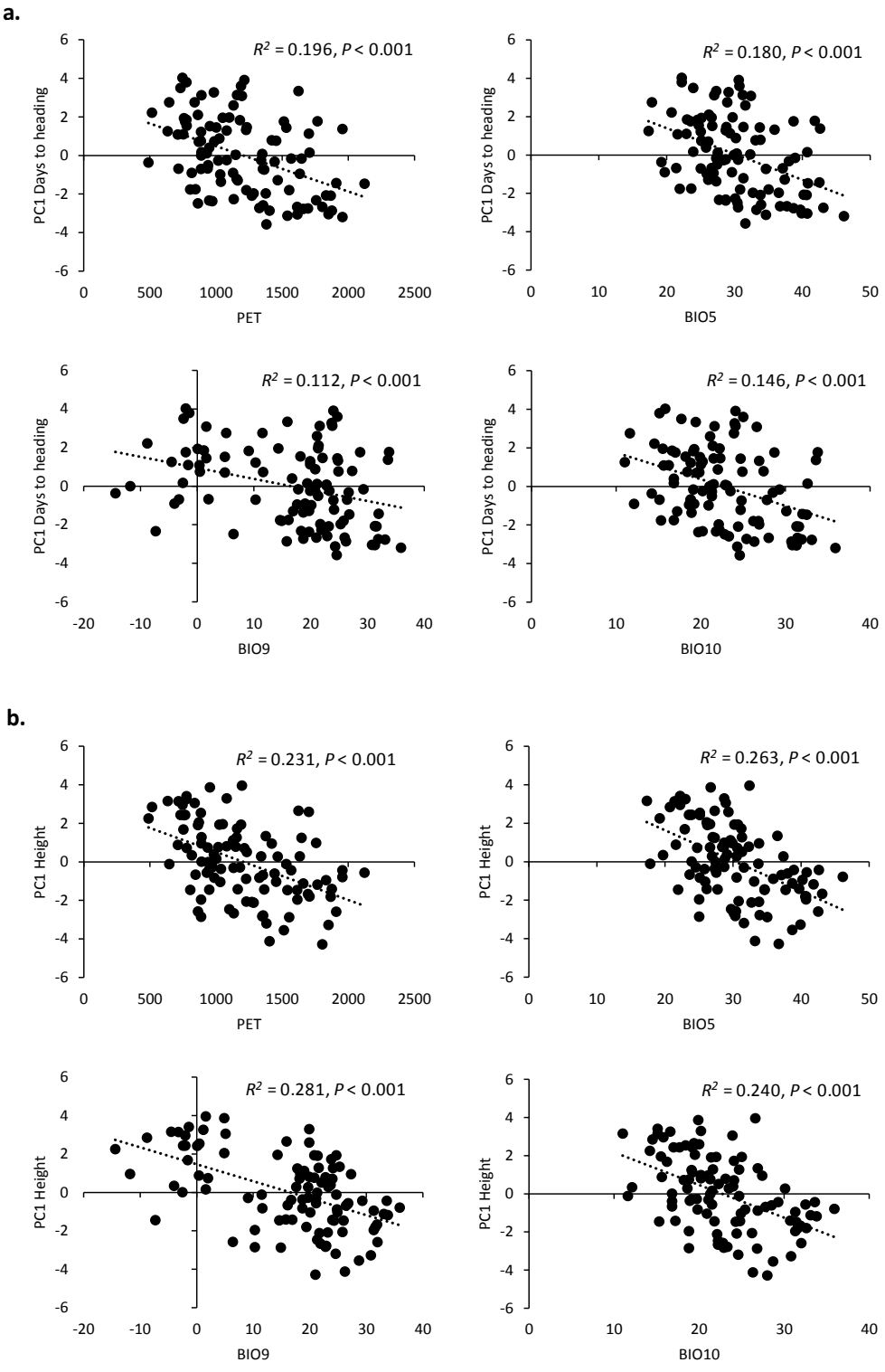


Supplementary Figure 6. Test for selective sweeps across geo-referenced spontaneum and landrace barley group accessions. Contigs with statistically significant values are circled. The statistically significant selective sweep position for the landrace group at 43.9 cM on 3H corresponds to a pentatricopeptide repeat (PPR) protein that is proximate to, and intermediate between, the positions of the important domestication genes *HvBTR1* and *HvBTR2* (*HvBTR2* at 40.7 cM) and the flowering-associated genes *HvFT2* (45.6 cM) and *HvGI* (45.8 cM). The corresponding bins indicated high reductions in diversity in landrace compared to spontaneum material (**Fig. 2a**).

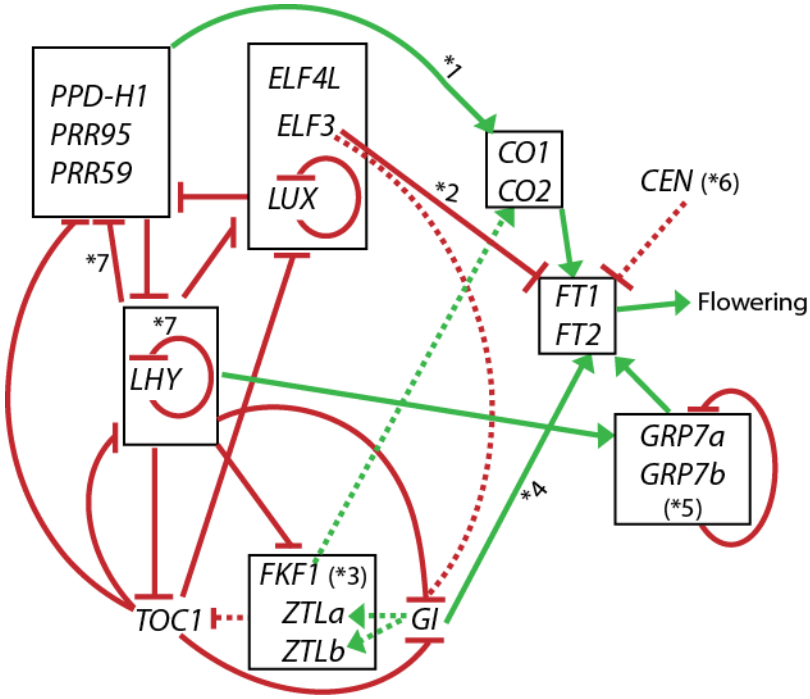
Supplementary Figure 7. X^TX and Bayes factor results. **a**. Pairwise correlations between 14 synthetic or individual environmental variables based on groupings assigned by the "pam" function in cluster and as used to calculate Bayes factors (**Online Methods**). **b**. The top 1% of Bayes factor SNPs for two Bayenv runs that were also in the top 5% of absolute correlations in each run for 15 environmental variables (variables as in **a**. but with the addition of latitude; see **Supplementary Table 3**). **c.** The top 0.1% of X^TX SNPs, with positions for different barley groups offset vertically for visualisation purposes (i.e., not scaled by absolute value). As in **b**., results are offset by barley group.



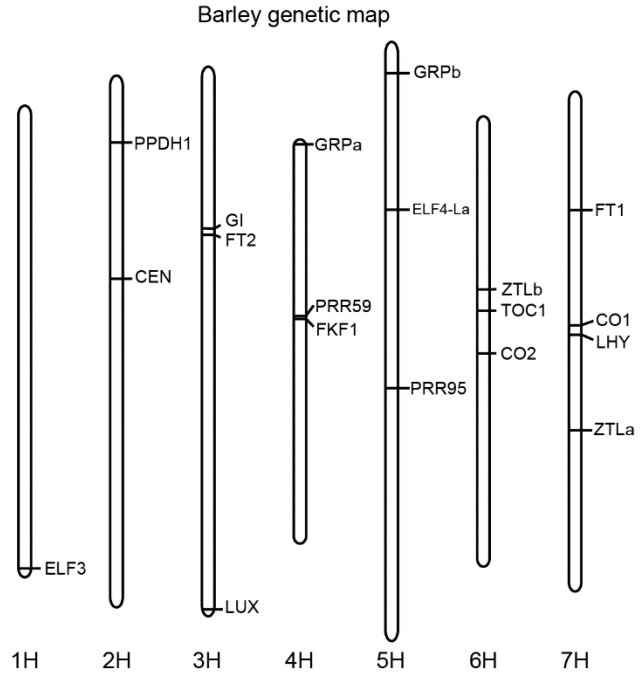




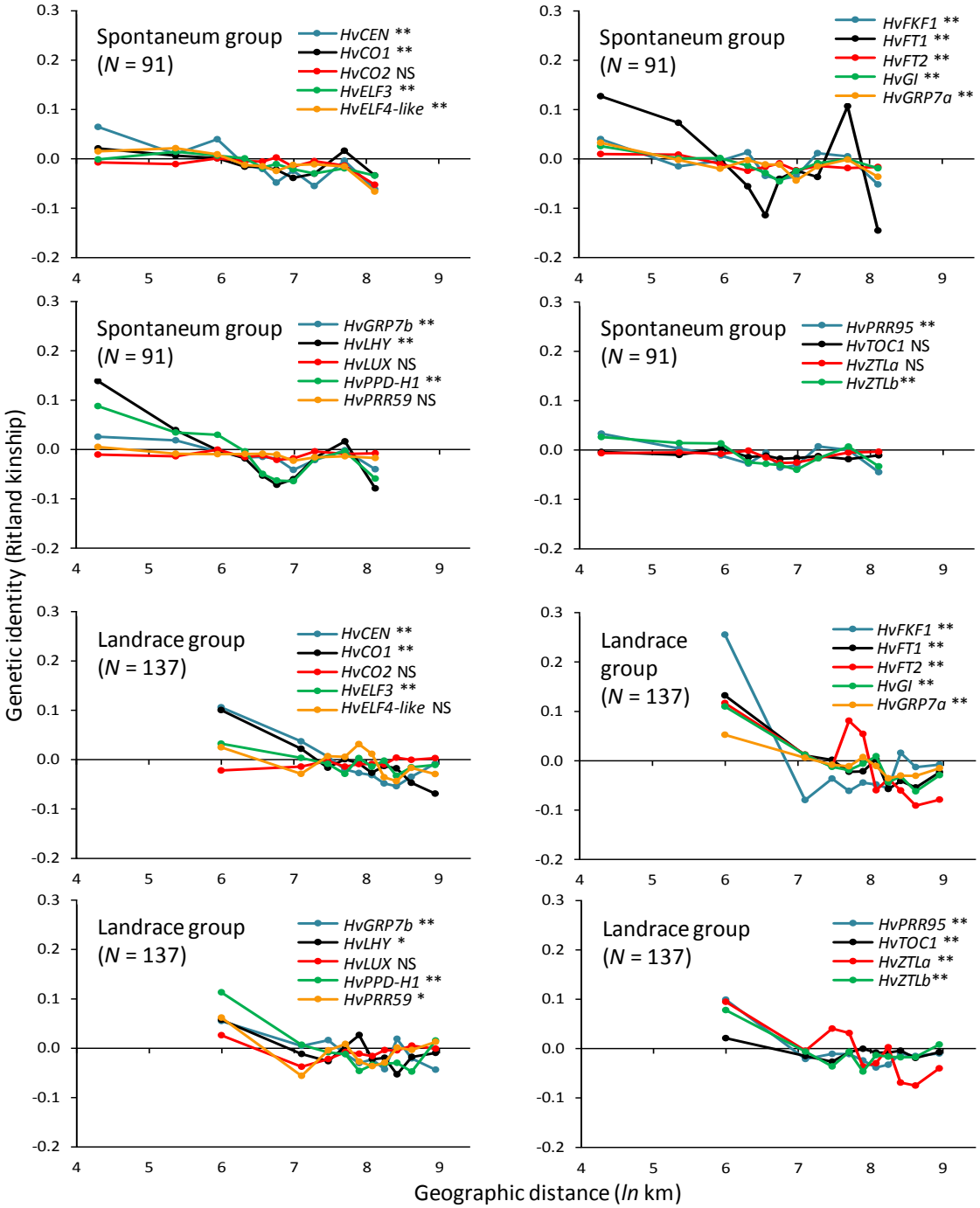
Supplementary Figure 8. Regression plots of individual components of the synthetic variable PET+BIO5+BIO9+BIO10 with PC1 days to heading **a.** and height **b**.



Supplementary Figure 9. Proposed schematic gene network of barley flowering-associated genes superimposed on the *Arabidopsis* model (after Calixto *et al*[1]). The gene regulations shown were compiled from Calixto *et al*[1] and individual articles marked by an * on the figure are referenced in the notes below. Full lines represent transcriptional feedback loops, whereas dashed lines represent post-translational regulation. Green lines are for activation, while red lines are for repression. Notes: *1 *PRR* genes are shown to activate transcription of *CO* genes but this is probably a de-repression process involving *CYCLING DOF FACTORS* (*CDFs*) as in *Arabidopsis*[2,3]. *2 In short days, *Arabidopsis* ELF3 represses transcription of *FT* genes through accumulation of the flowering repressor SHORT VEGETATIVE PHASE (SVP)[4]. *3 In *Arabidopsis,* the CO protein is stabilised by several components in daytime, including FKF1[5]. Additionally, FKF1 interacts with GI and regulates *CO* expression in long days[6]. *4 AtGI directly activates the expression of *AtFT*[7]. *5 In *Arabidopsis*, *GRP7* is probably regulated by CCA1/LHY[8] and it also auto-regulates its expression[9]. Here we show GRP7a/b directly activating transcription of *FTs* but this is probably a de-repression process involving FLOWERING LOCUS C (FLC) as in *Arabidopsis*[10]. *6 *Arabidopsis* CENTRORADIALIS homologue (ATC) inhibits flowering probably by antagonizing FT activity[11]. *7 Newly uncovered negative auto-regulatory feedback loops in *Arabidopsis* from *LHY* and the *PRRs* genes[12].



Supplementary Figure 10. Distribution of 19 flowering-associated genes used in our study across chromosomes on the barley genome.

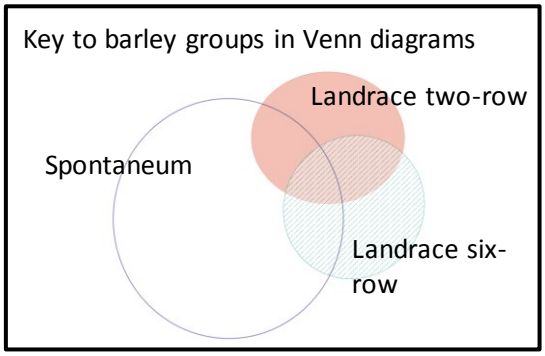


Supplementary Figure 11. Spatial autocorrelation analyses of genic SNPs for flowering-associated genes separately (up to five genic profiles shown on each graph to aid visualisation; *P* values for overall structure shown for each profile, NS = not significant, * = $P < 0.05$, ** = $P < 0.01$). Levels of geographic structuring vary by gene and barley group, as also observed in **Supplementary Fig. 12-30**. The high structuring for *HvFKF1* (a member of the LOV blue light receptor subfamily and a core clock component) in the landrace group is due to comparatively rare alleles at a number of SNPs that are in high linkage disequilibrium occurring in geographically proximate accessions. For *HvTOC1* and *HvZTLa*, geographic structuring was statistically insignificant in the spontaneum group but was significant in landrace material, which may indicate specific domestication processes (see also **Supplementary Fig. 12-30**). *HvLUX* contains high genetic variation in the spontaneum group, but diversity is not statistically significantly geographically structured in either barley group.

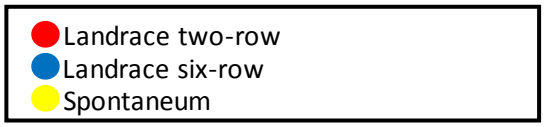
Detailed legend descriptions for Supplementary Figure 12 - 30. Information on haplotype profiles for 19 flowering-associated genes in spontaneum and landrace groups. For each gene, the number of SNP positions scored to construct haplotypes (*n*) is indicated. Each figure shows **a.** The individually curated structures of each gene. Non-synonymous SNPs are represented by large red arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except for conserved domain-encoding exons, which are colored. **b.** (landraces) and **c.** (spontaneum) show the geographic distribution of the observed haplotypes ordered from most frequent haplotype (red, A) to grouped unique or rare haplotypes (orange,I). All accessions shown in each map have complete data at each gene. **d.** Venn diagrams for haplotypes, in this case sub-dividing the landrace group into two- and six-row inflorescence types. For each gene, the total number of haplotypes revealed in the whole collection (*H*) is indicated. **e.** Median-joining networks for each flowering-associated gene showing the relationship between spontaneum and landrace haplotypes, again structuring landrace into two- and six-row types. The area of a circle is proportion to the number of accessions with a given haplotype. All metrics indicate a higher (generally, much higher) genetic diversity in the spontaneum group than in landraces for each of the 19 tested genes. The keys required to interpret information in **b., c., d.,** and **e.** of each figure are as follows (also shown for the first gene).

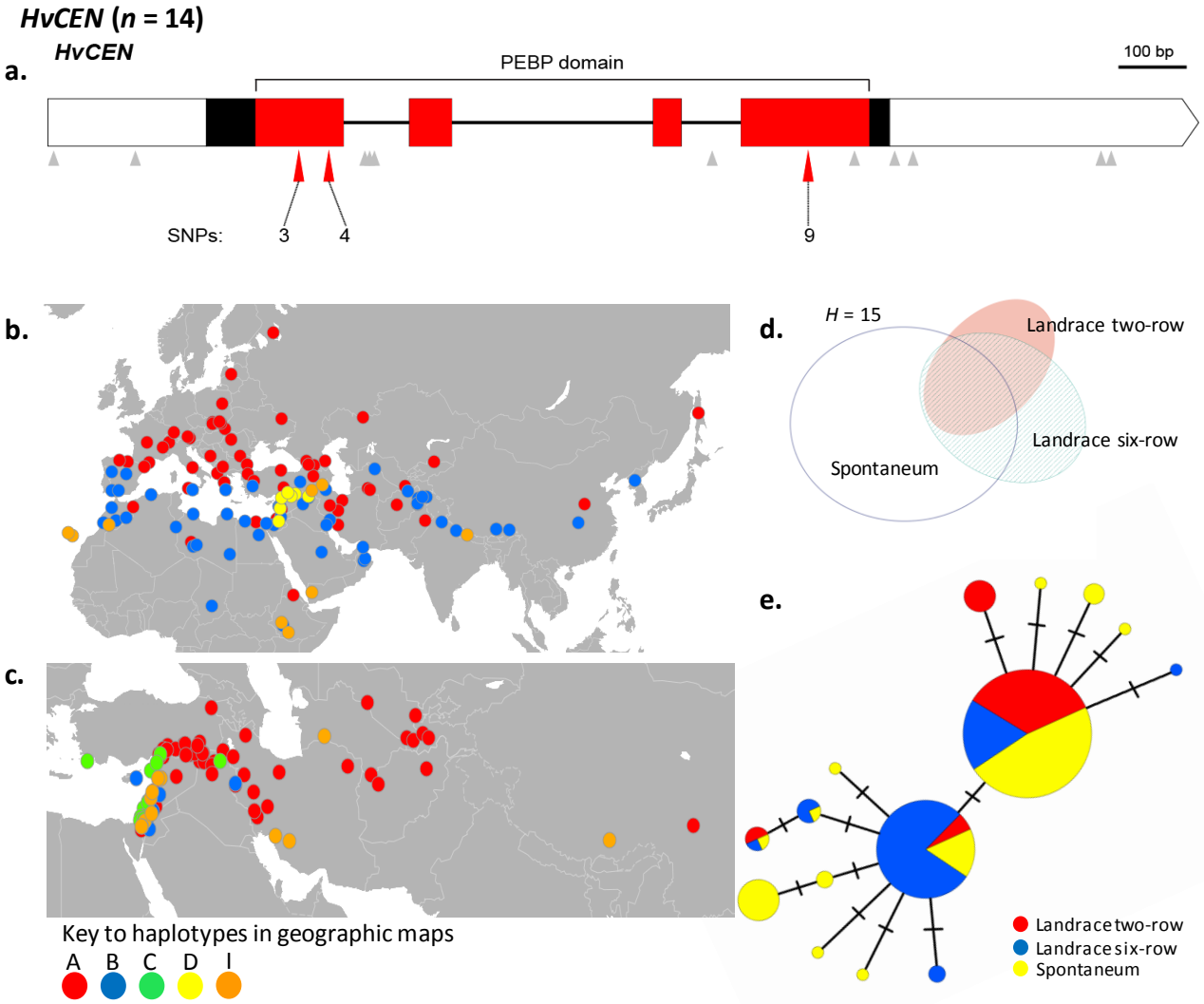
b. and **c.**

d.

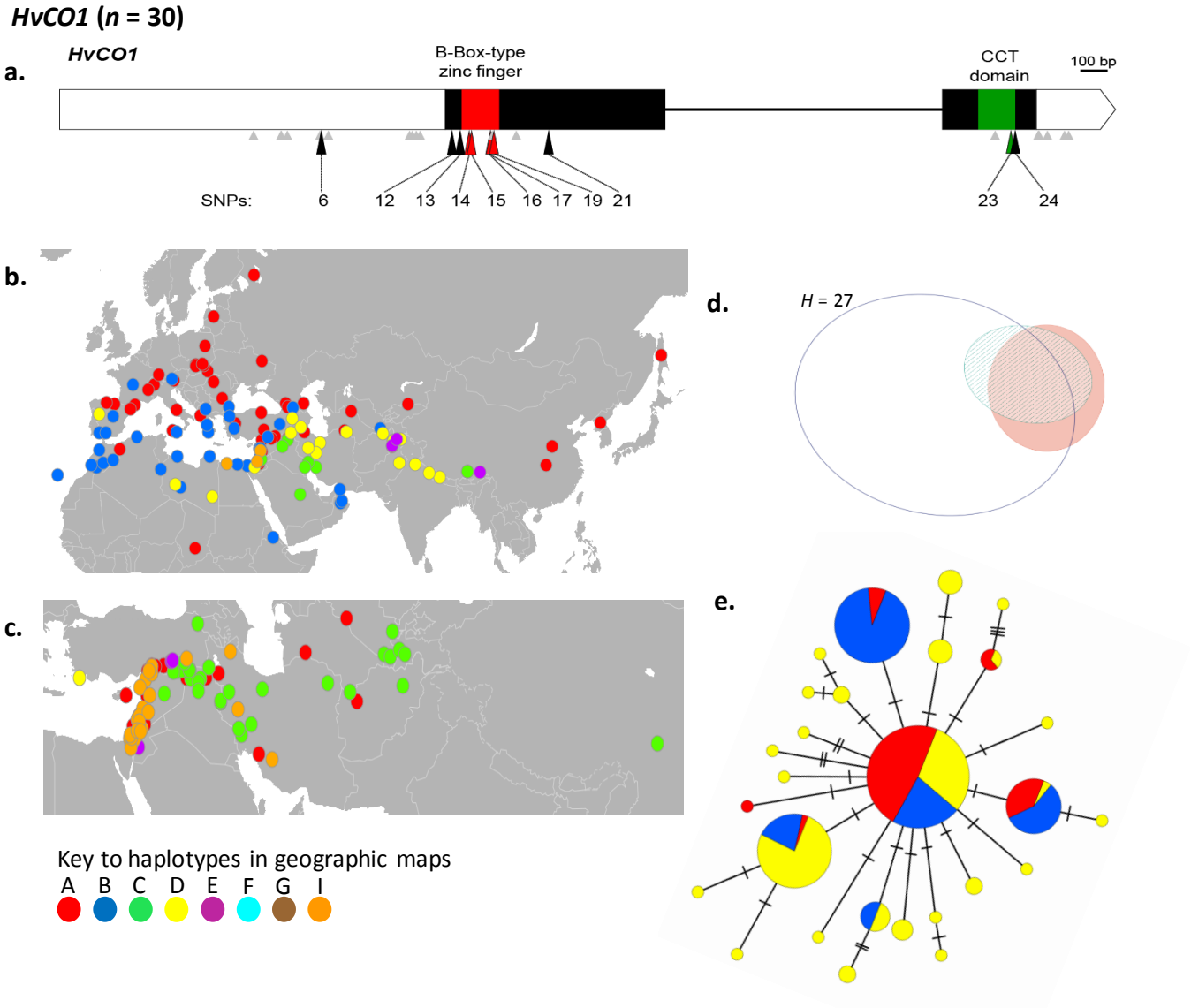


e.

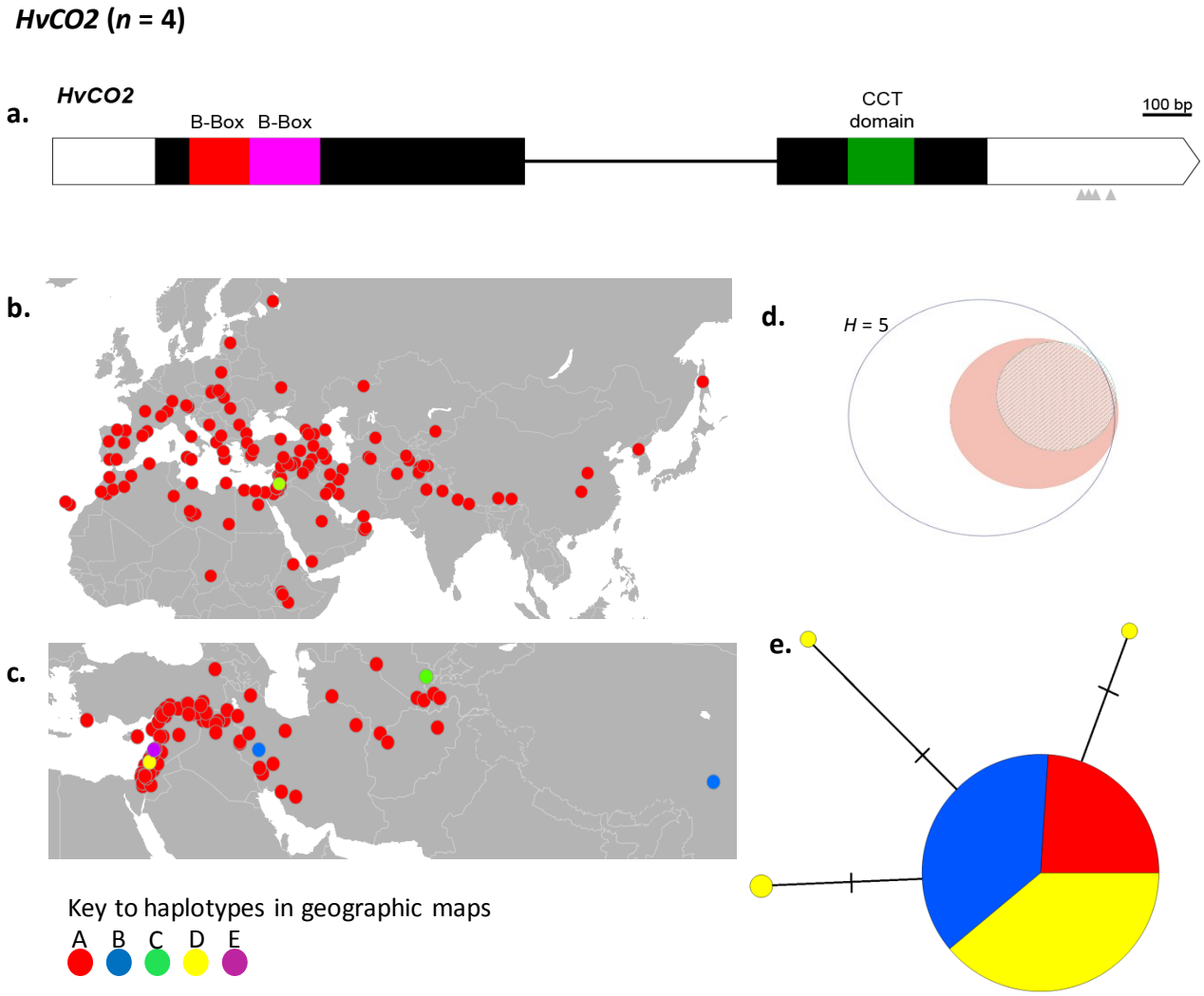




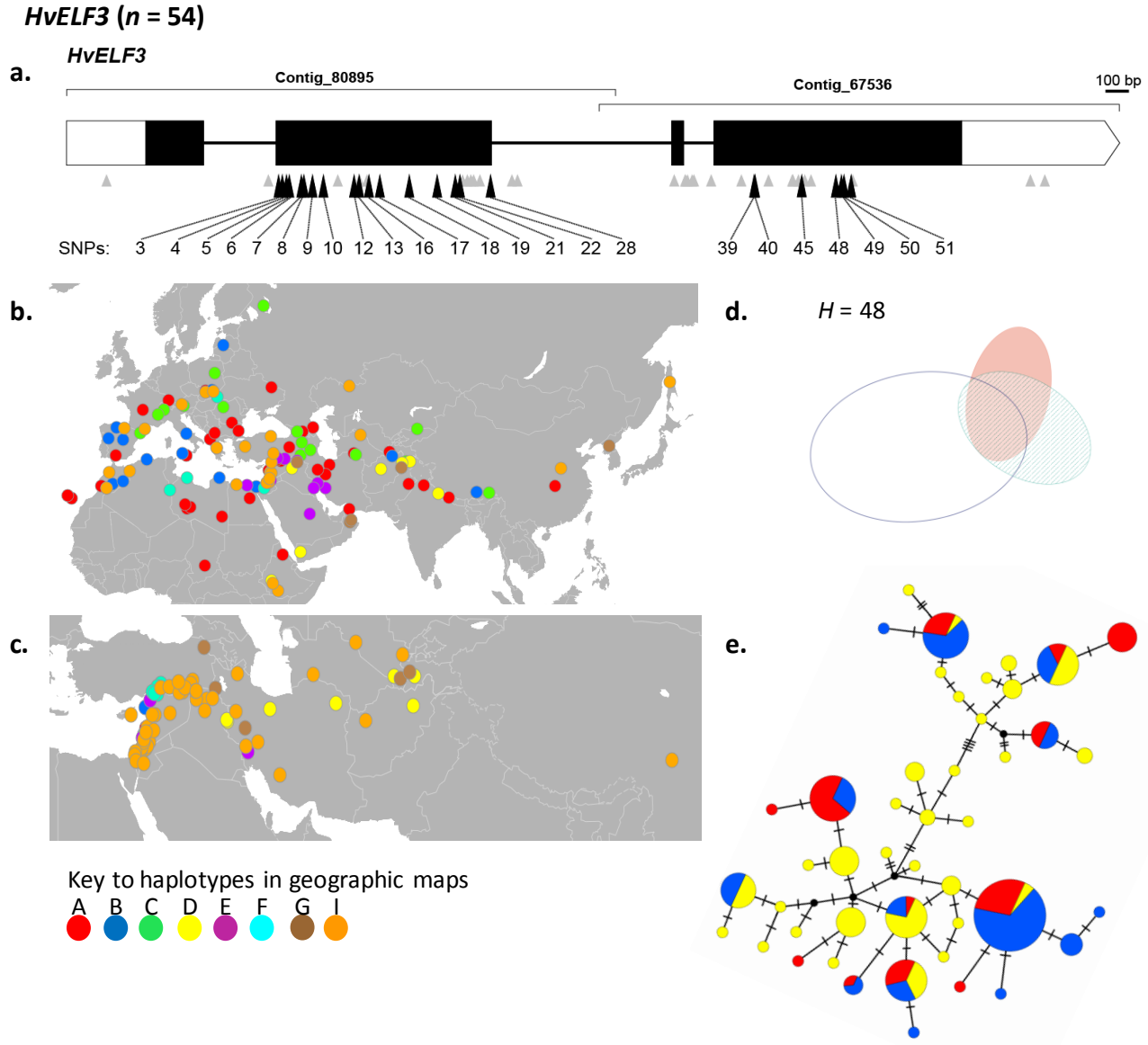
Supplementary Fig. 12: **a.** Genomic structure of *HvCEN* and relative positions and effect of the 14 SNPs identified. Synonymous SNPs and SNPs in UTRs and introns are indicated by small grey arrows. Non-synonymous SNPs are represented by large red arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (PhosphatidylEthanolamine-Binding Protein (PEBP) domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups, **e.** Median joining network. Further details given in general legend.



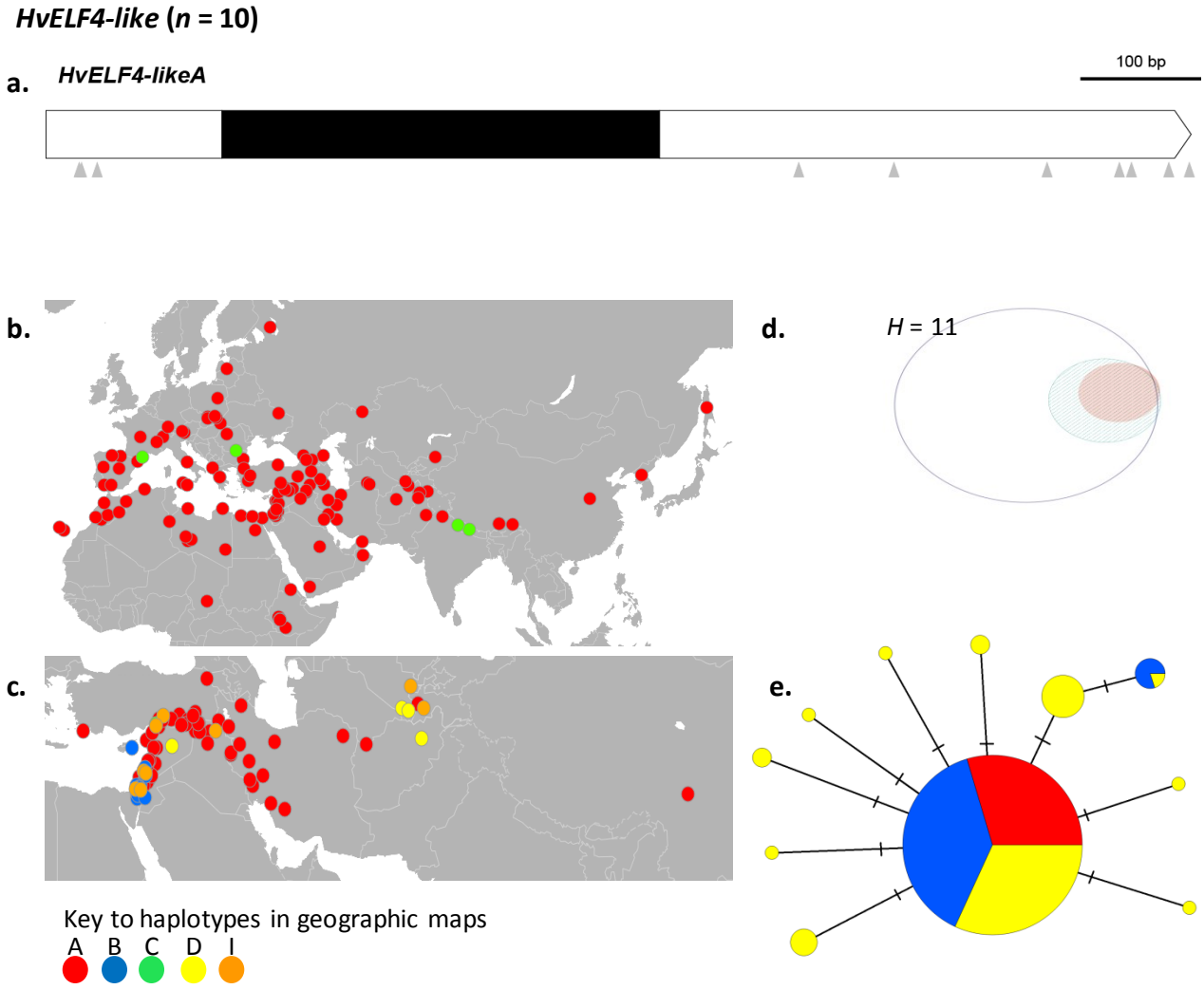
Supplementary Fig. 13: **a.** Genomic structure of *HvCO1* and positions of the 30 SNPs identified. Synonymous SNPs and SNPs in UTRs are indicated by small grey arrows. SNP 6, which shortened uORF from 38 to 15 amino acids and non-synonymous SNPs are represented by large colour coded arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (B-Box-type zinc finger) or green (CCT domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups, **e.** Median joining network. Further details given in general legend.



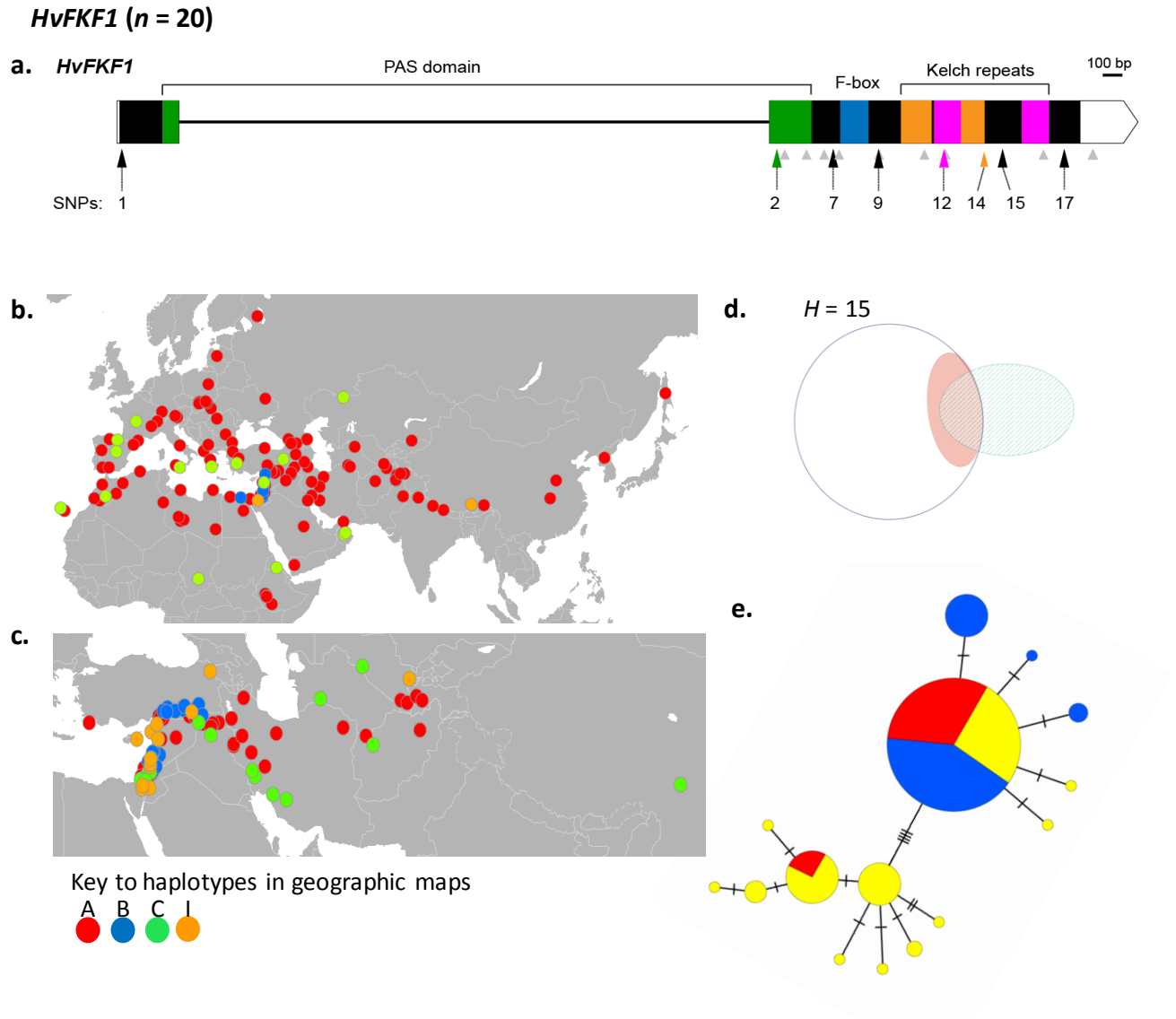
Supplementary Fig. 14: a. Genomic structure of *HvCO2* and positions of the 4 SNPs identified. SNPs are in the 3' UTR and are indicated by small grey arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red/pink (B-Box-type zinc finger) or green (CCT domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups, **e.** Median joining network. Further details given in general legend.



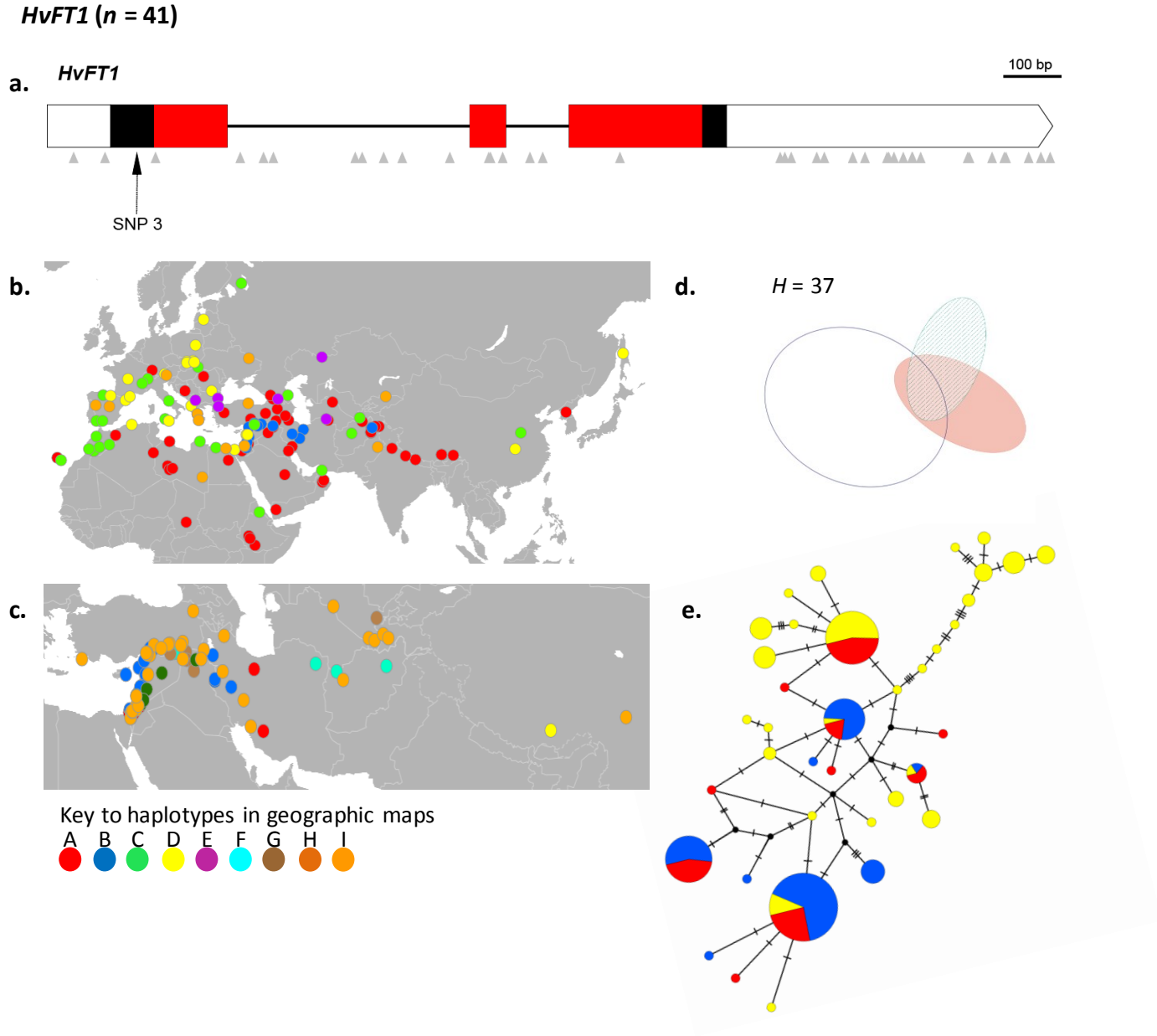
Supplementary Fig. 15: **a.** Genomic structure of *HvELF3* and positions of the 54 SNPs identified. Synonymous SNPs and SNPs in UTRs are indicated by small grey arrows. Non-synonymous SNPs are represented by large black arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



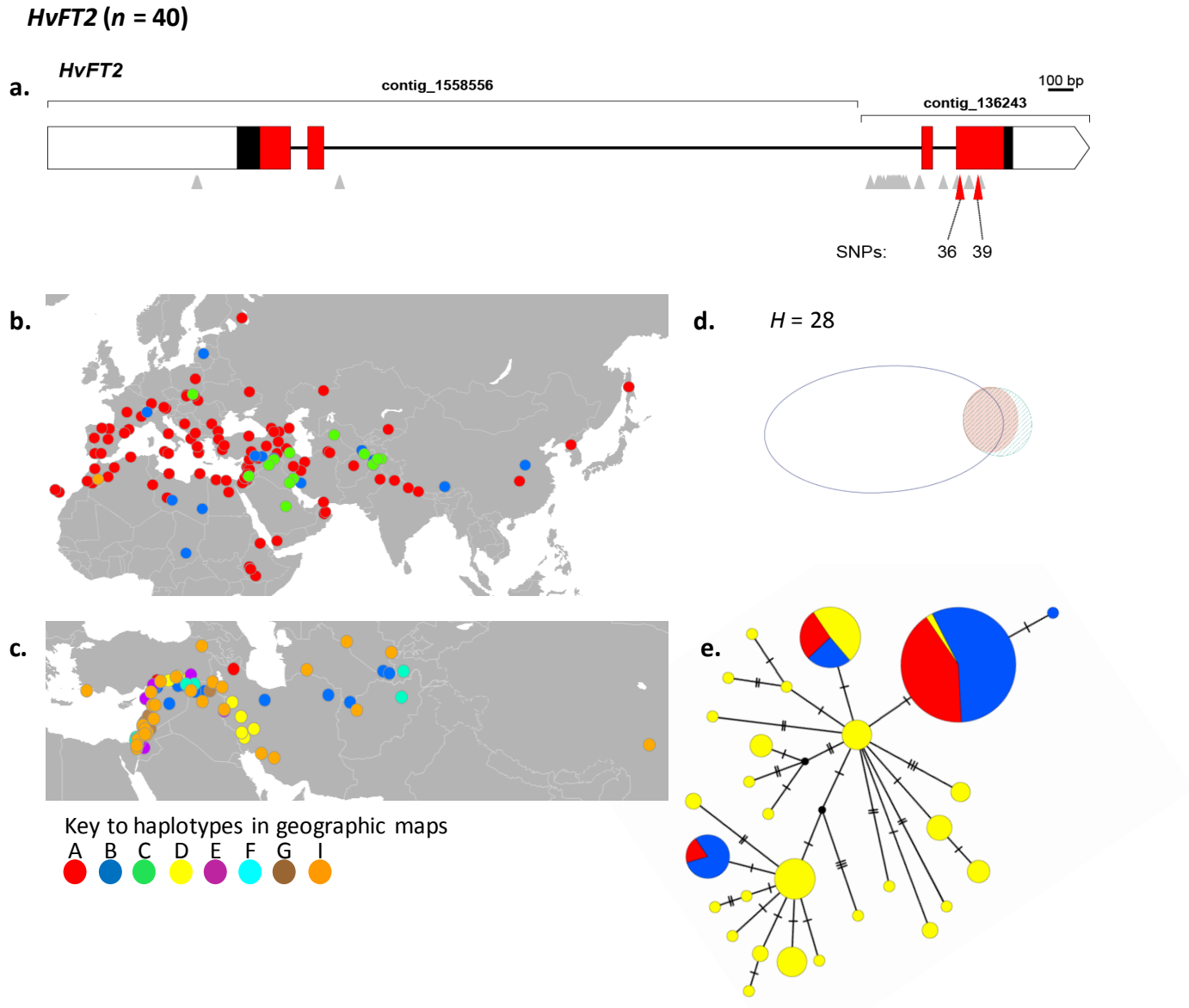
Supplementary Fig. 16: **a.** Genomic structure of *HvELF4-LikeA* and positions of the 10 SNPs identified. SNPs are indicated by small grey arrows. 5' and 3' UTRs are open boxes; coding sequence is the dark box **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



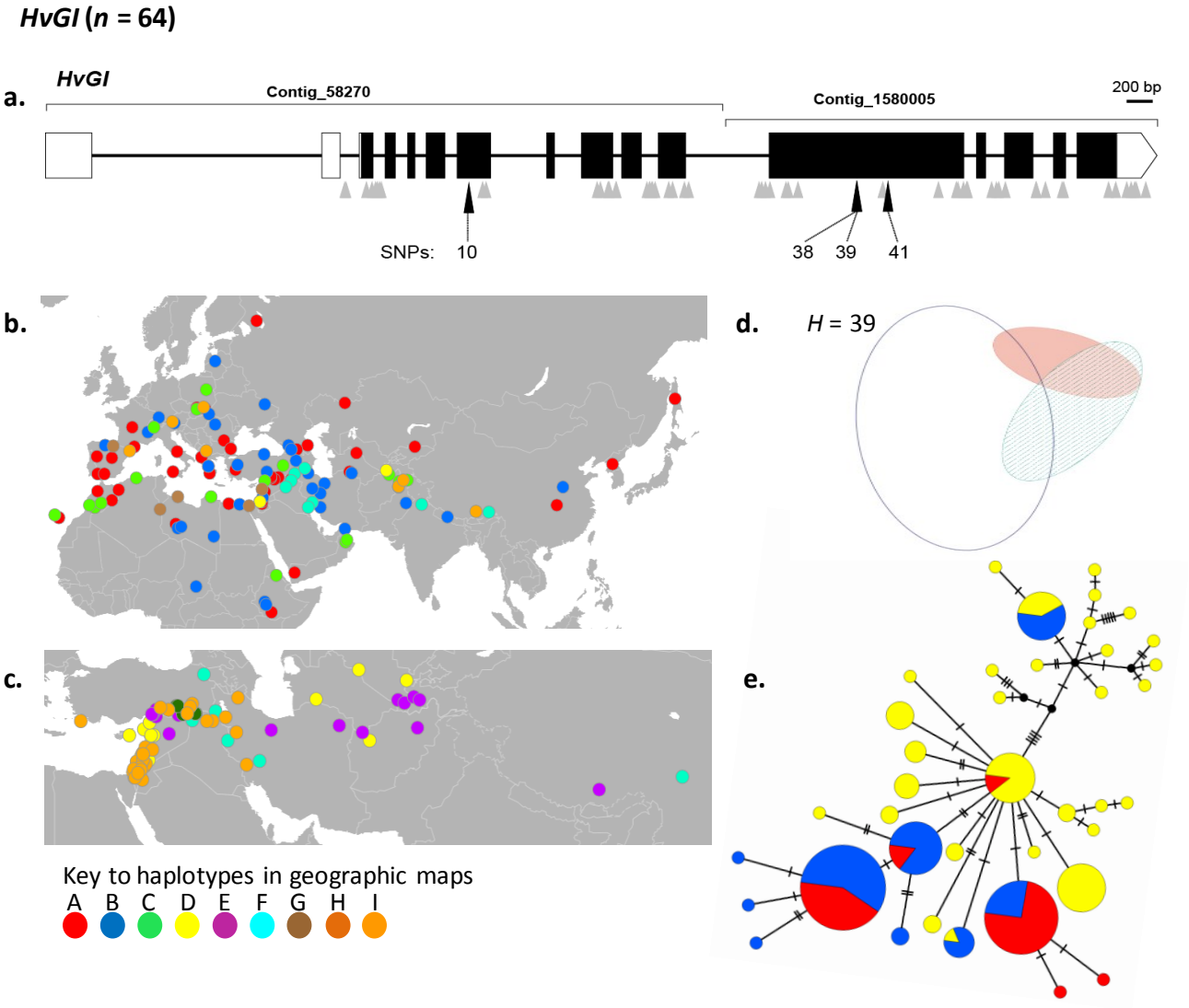
Supplementary Fig. 17: **a.** Genomic structure of *HvFKF1* and positions of the 20 SNPs identified. Synonymous SNPs and SNPs in UTRs are indicated by small grey arrows. Non-synonymous SNPs are represented by large colour coded arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded green (PAS domain), blue (F-box-type domain), or orange/pink (Kelch motifs) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



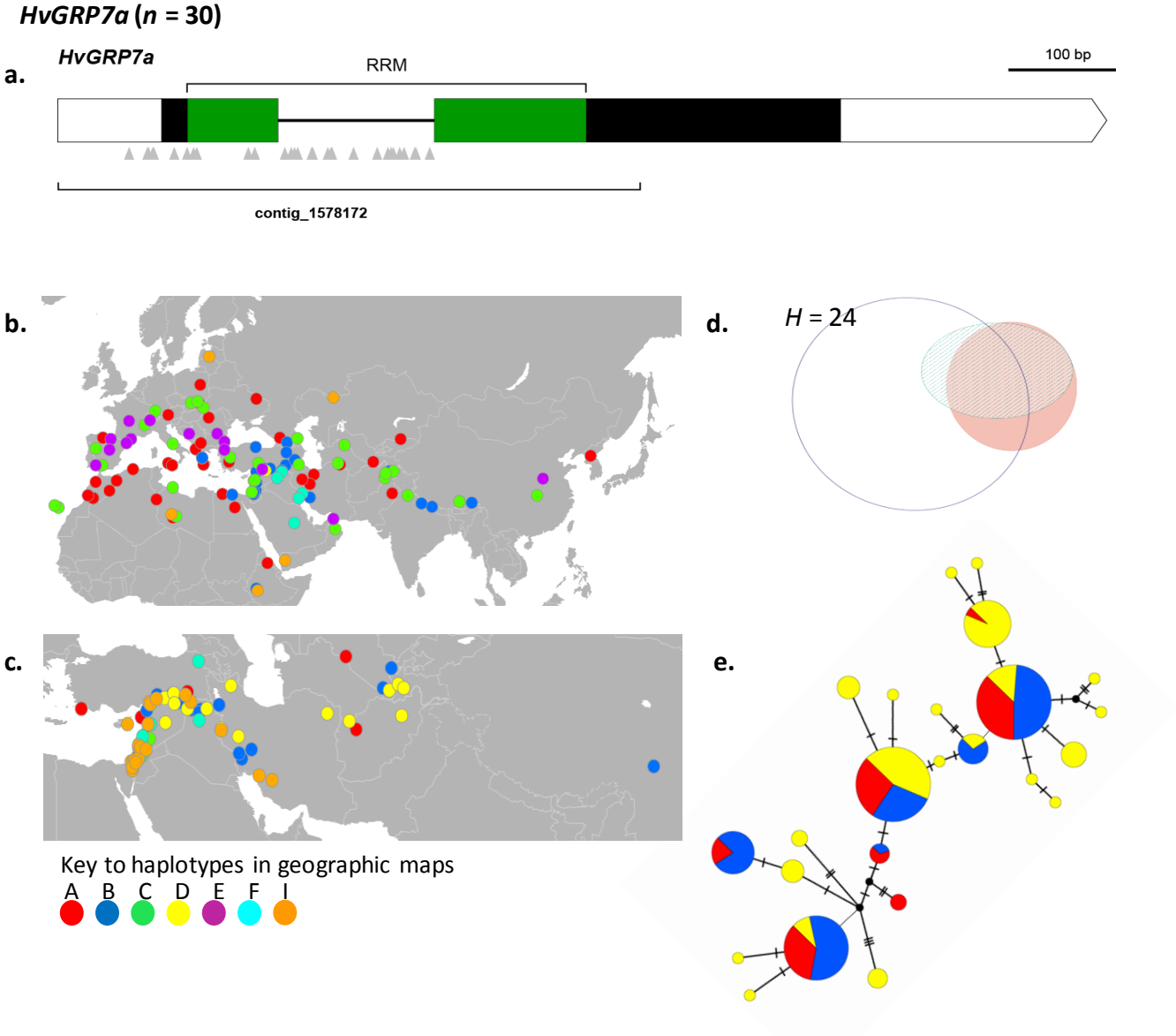
Supplementary Fig. 18: **a.** Genomic structure of *HvFT1* and positions of the 41 SNPs identified. Synonymous SNPs and SNPs in UTRs and introns are indicated by small grey arrows. Non-synonymous SNP 3 is represented by a large black arrow. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (phosphatidylethanolamine-binding protein domain (PEBP)) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



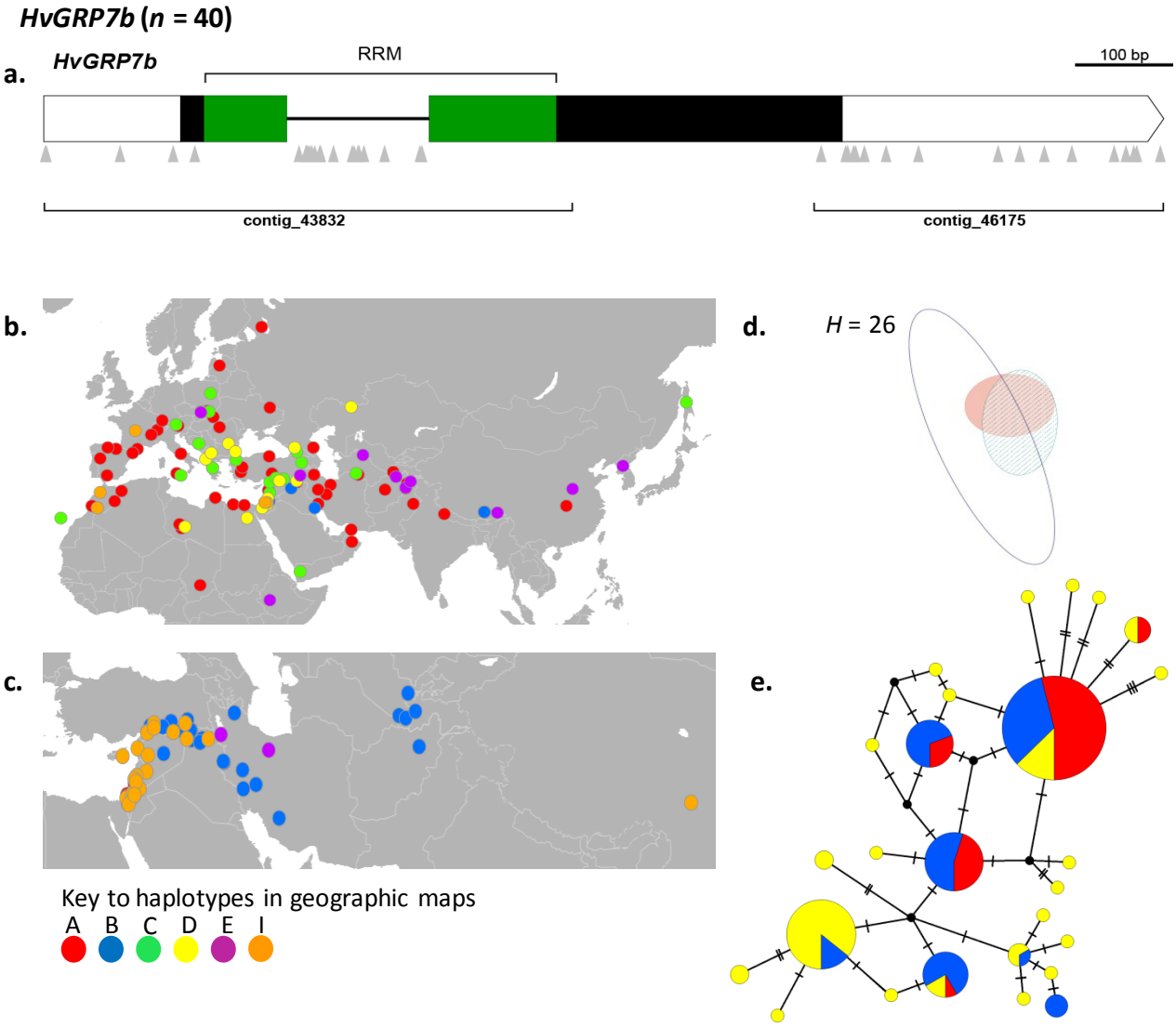
Supplementary Fig. 19: **a.** Genomic structure of *HvFT2* and positions of the 40 SNPs identified. Synonymous SNPs and SNPs in 5'UTR and introns are indicated by small grey arrows. Non-synonymous SNPs are represented by large red arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (phosphatidylethanolamine-binding protein domain (PEBP)) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



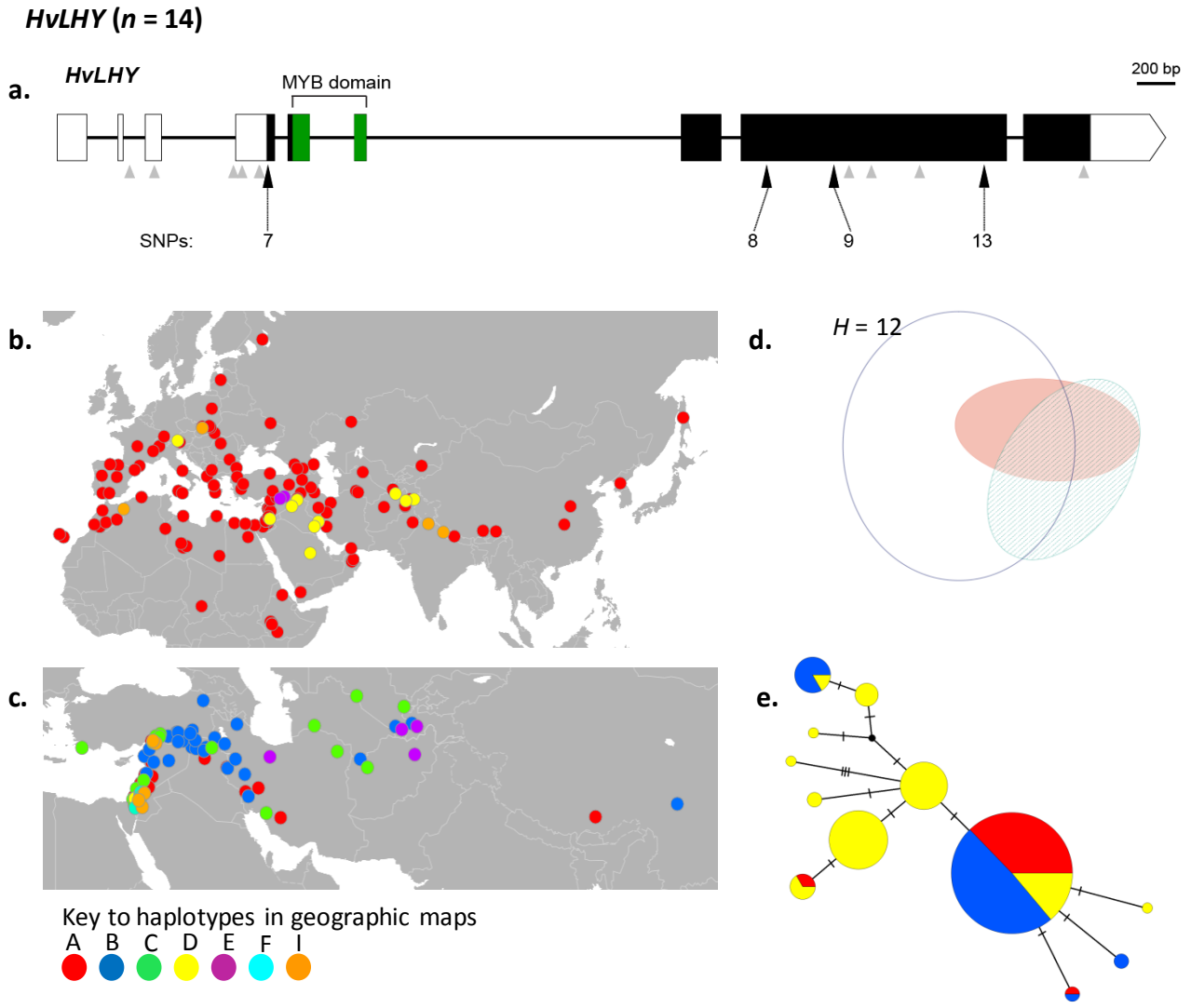
Supplementary Fig. 20: **a.** Genomic structure of *HvGI* and positions of the 64 SNPs identified. Synonymous SNPs and SNPs in UTRs and introns are indicated by small grey arrows. Non-synonymous SNPs are represented by large black arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



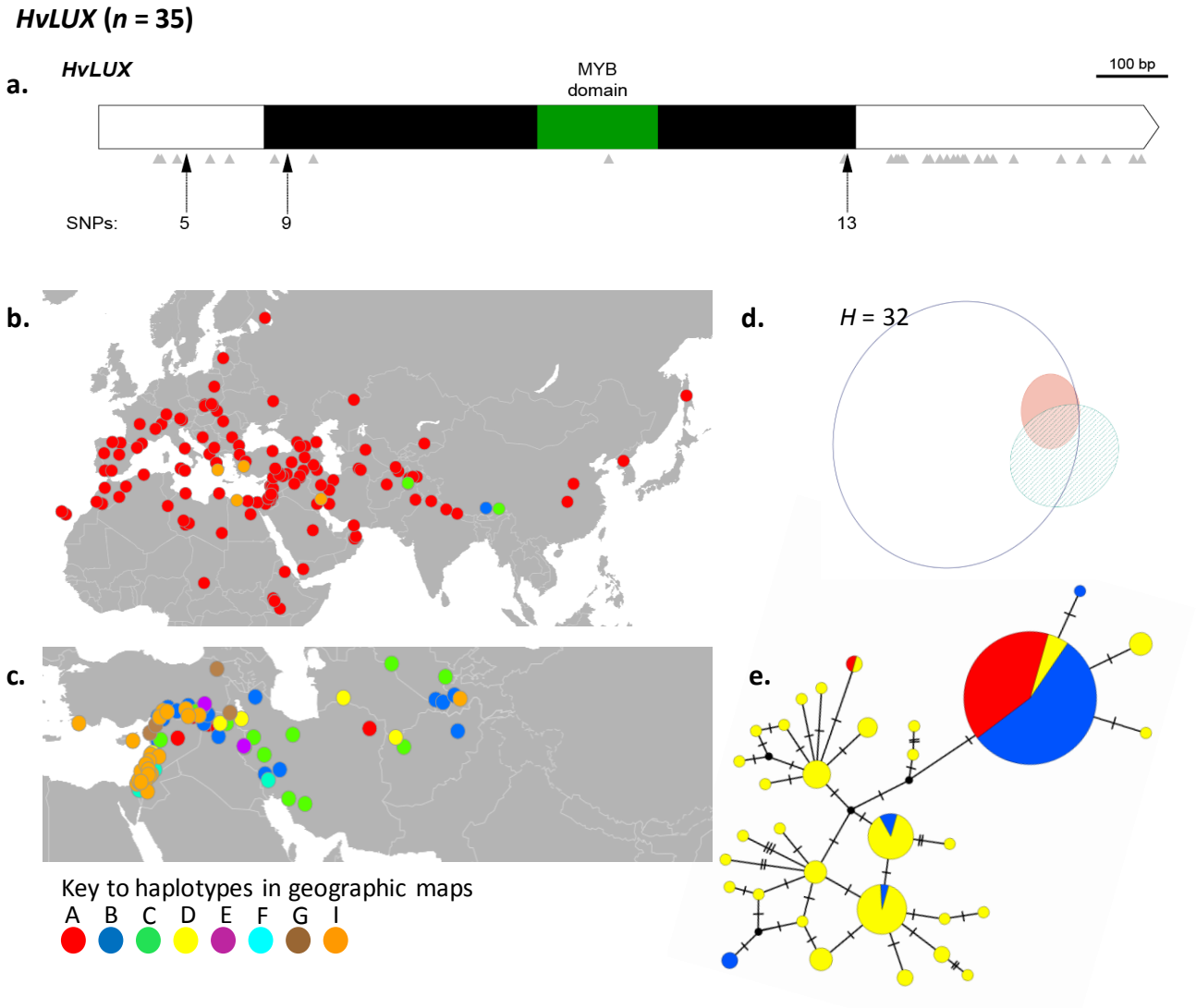
Supplementary Fig. 21. **a** Putative genomic structure of *HvGRP7a* and positions of the 30 SNPs identified. Synonymous SNPs, and SNPs in the 5'UTR and intron are indicated by small grey arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded green (Green - RNA Recognition motif: RRM) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



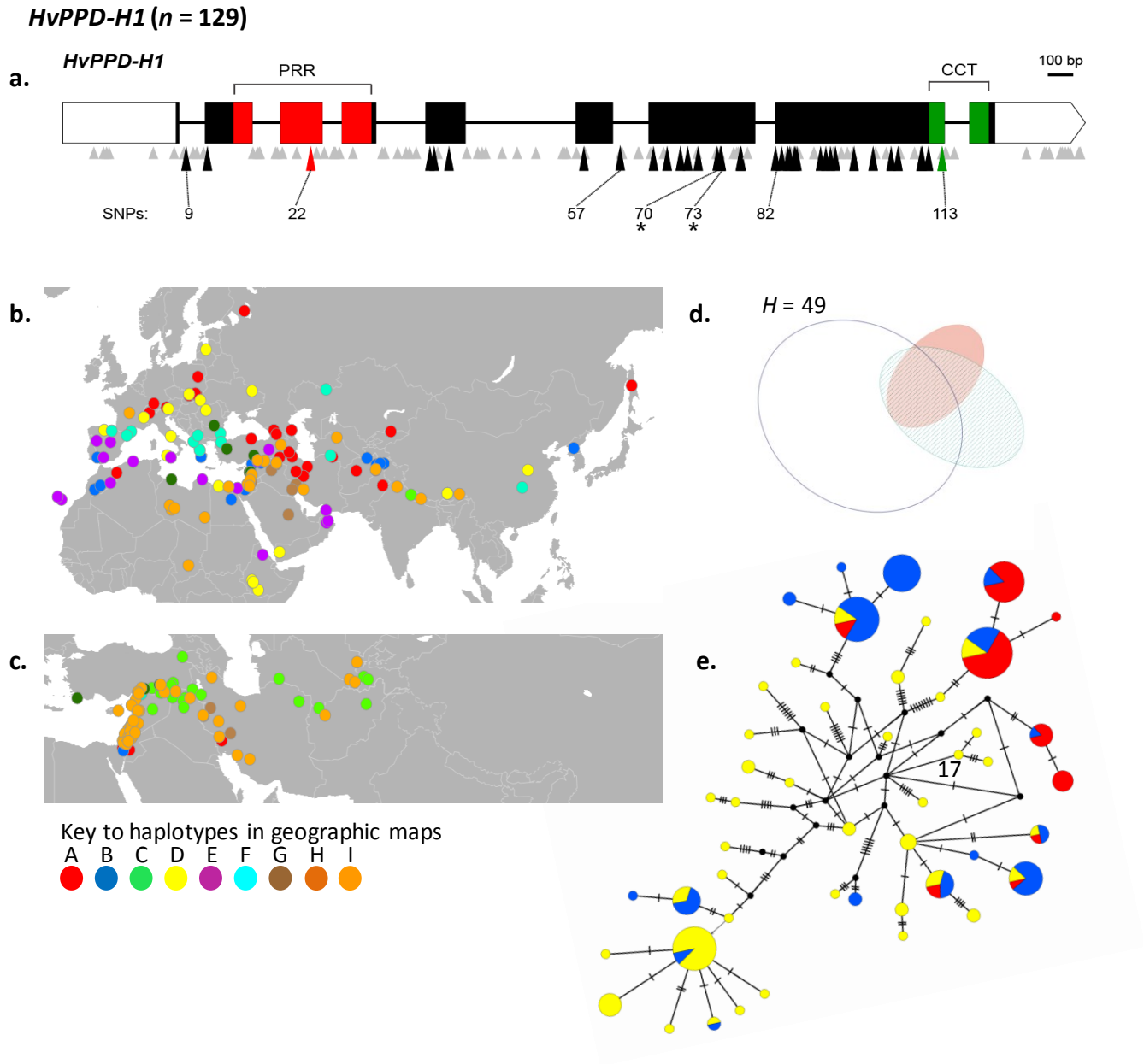
Supplementary Fig. 22: a. Genomic structure of *HvGRP7b* and positions of the 40 SNPs identified. Synonymous SNPs and SNPs in UTRs and intron are indicated by small grey arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded green (Green - RNA recognition motif: RRM) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



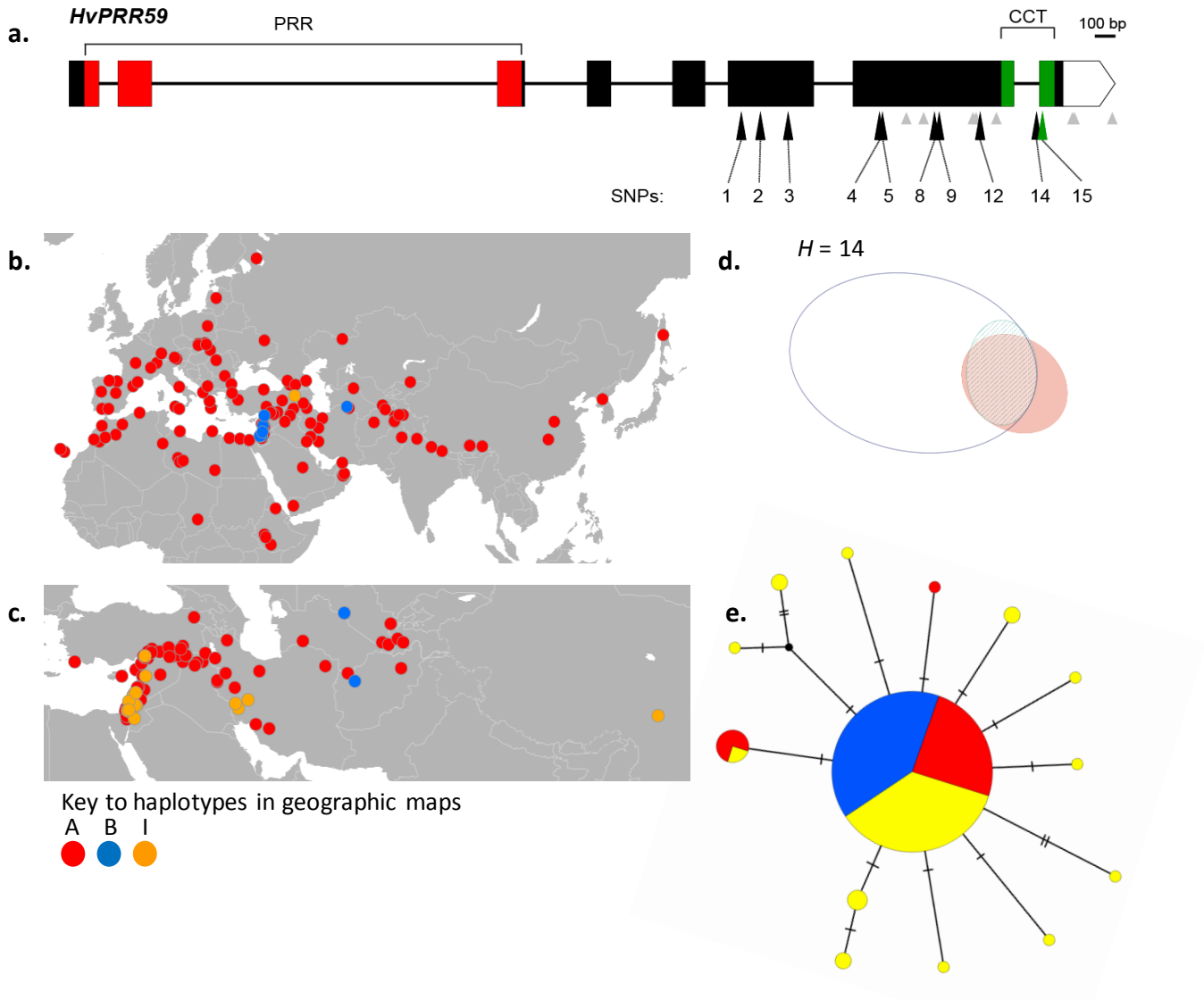
Supplementary Fig. 23: **a.** Genomic structure of *HvLHY* and positions of the 14 SNPs identified. Synonymous SNPs and SNPs in UTRs are indicated by small grey arrows. Non-synonymous SNPs are represented by large black arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded green (MYB-like DNA binding domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



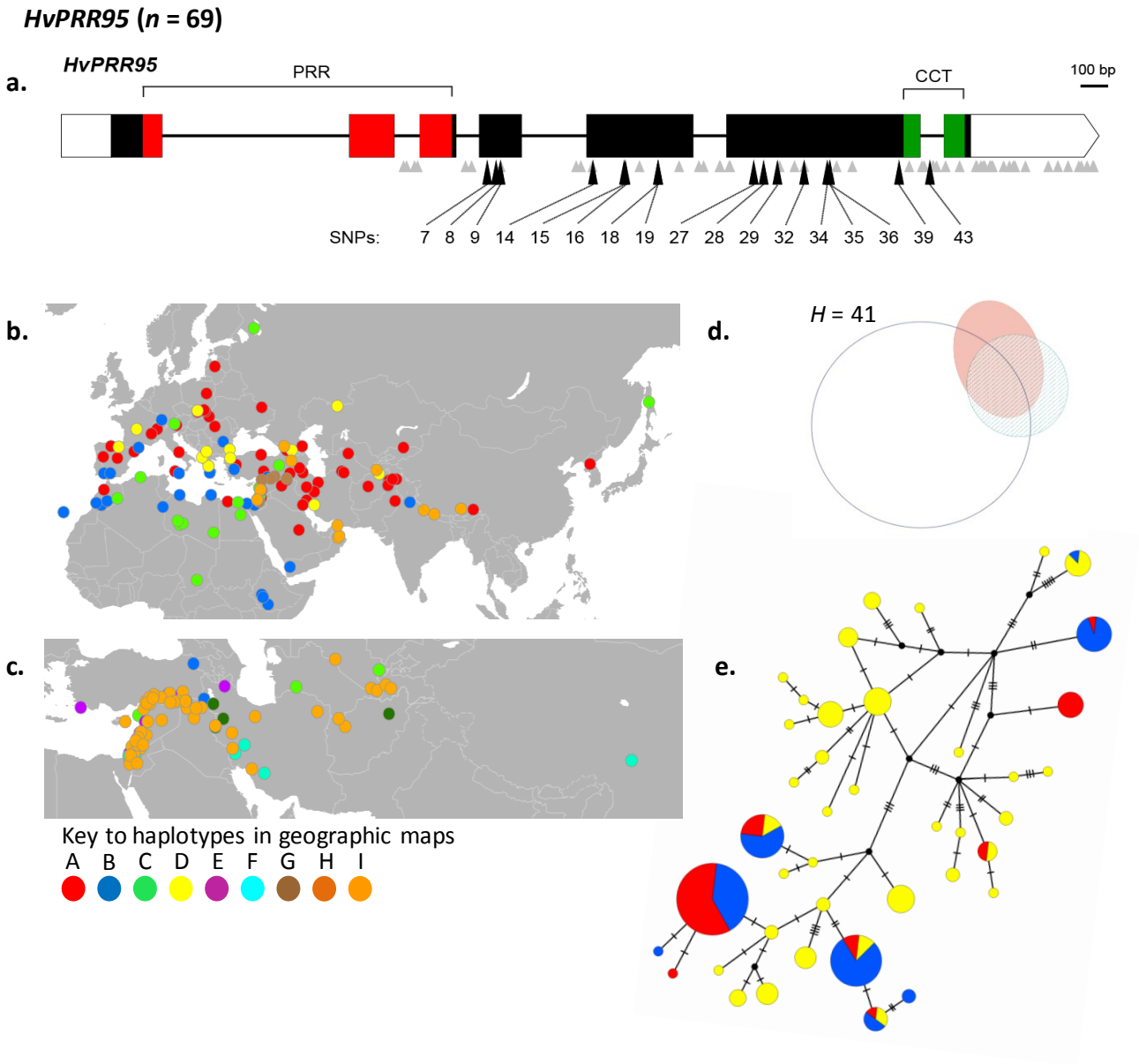
Supplementary Fig. 24: **a.** Genomic structure of *HvLUX* and positions of the 35 SNPs identified. Synonymous SNPs and SNPs in UTRs are indicated by small grey arrows. SNP 5, which introduced an uAUG (creating a short uORF of 7 amino acids) and non-synonymous SNPs are represented by large black arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding region, which is shaded green (MYB-like DNA binding domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



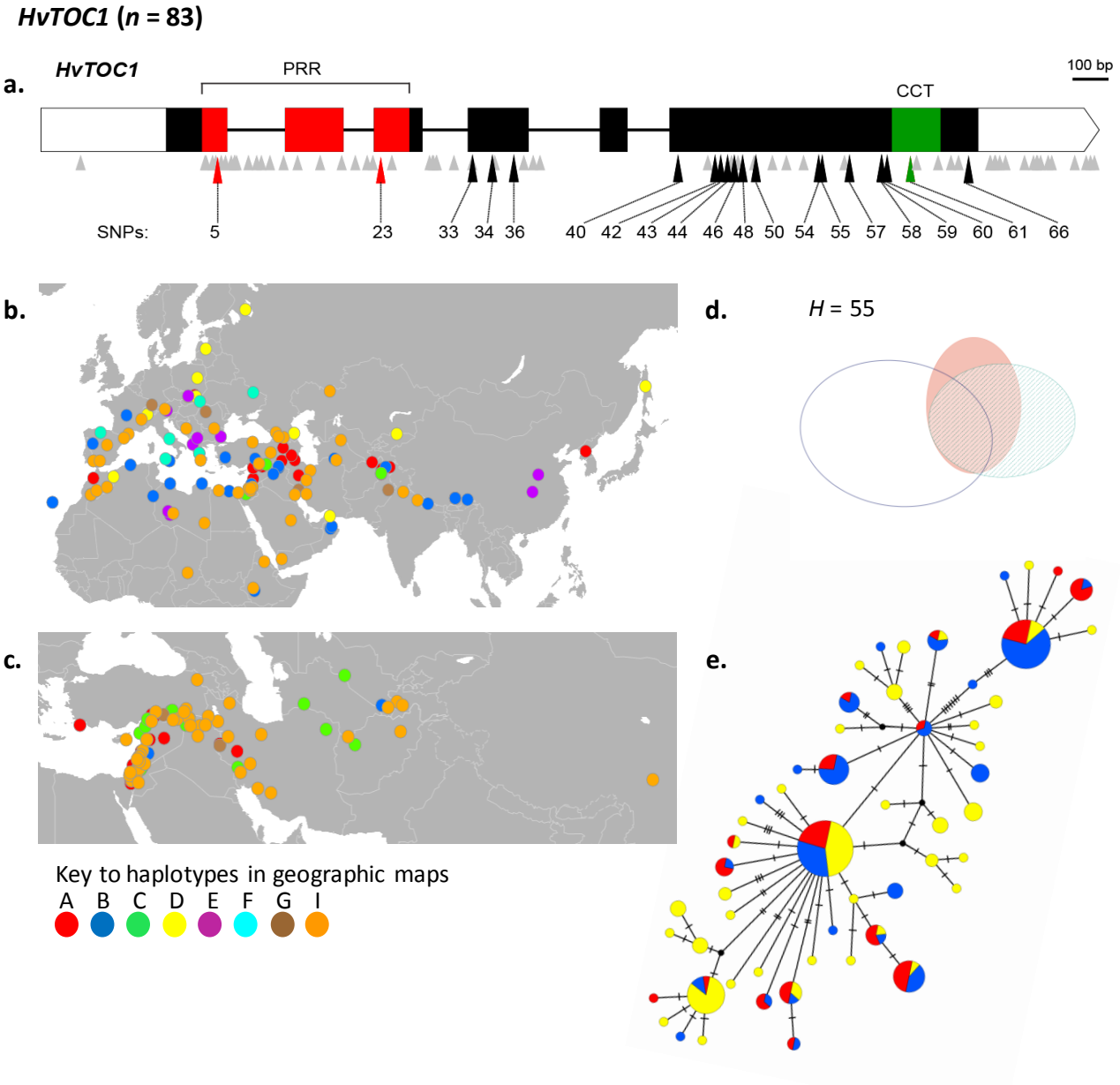
Supplementary Fig. 25: a. Genomic structure of *HvPPDH1* and positions of the 129 SNPs identified. Synonymous SNPs and SNPs in the 3'UTR and introns are indicated by small grey arrows. SNP 9, SNP 57, SNP 82 and non-synonymous SNPs are represented by large black arrows. SNP 9 reduces uORF of transcripts retaining Intron1 from 45 amino acids, which overlaps functional ORF, to 12 amino acids, which do not overlap functional ORF. SNP 57 is non-synonymous in transcripts undergoing common alternative 5' splice site of Exon 6 that adds 45 nt (Calixto pers. comm). SNP 82 changes 3' splice site from canonical AG to non-canonical AC. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (Pseudo-response regulator) or green (CCT domain). * SNPs 70 and 73 are associated with *Ppd-H1* or *ppd-H1* alleles according to Turner *et al.*[13] and Jones *et al.*[14], respectively **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



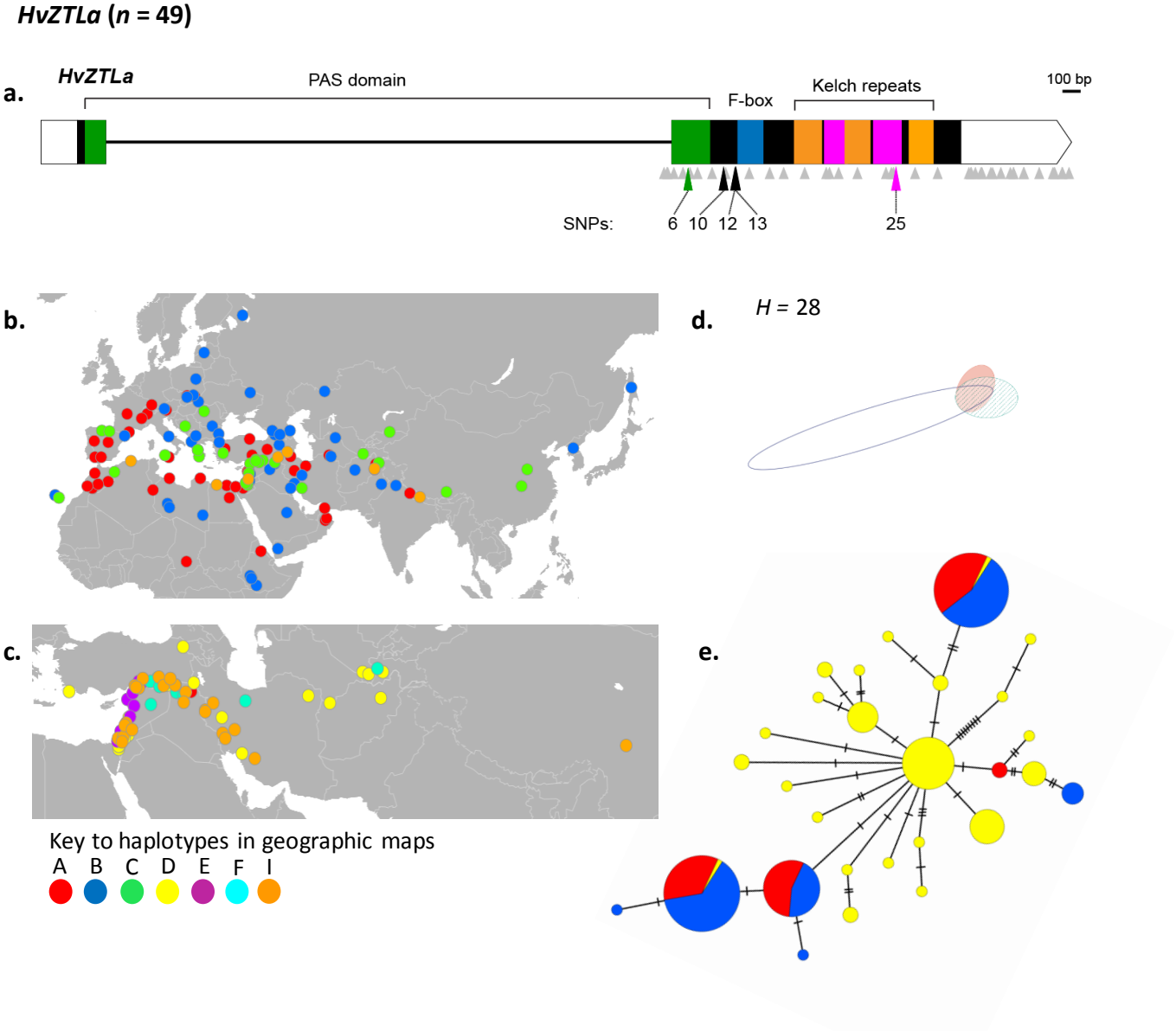
Supplementary Fig. 26: **a.** Genomic structure of *HvPRR59* present and positions of the 18 SNPs identified. Synonymous SNPs and SNPs in the 3'UTR are indicated by small grey arrows. SNP 14, which is a non-synonymous SNP for I7R transcripts (Intron7 has no PTC and is in frame), and non-synonymous SNPs are represented by large colour coded arrows. 3' UTR is the open box; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (Pseudo-Response Regulator) or green (CCT domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



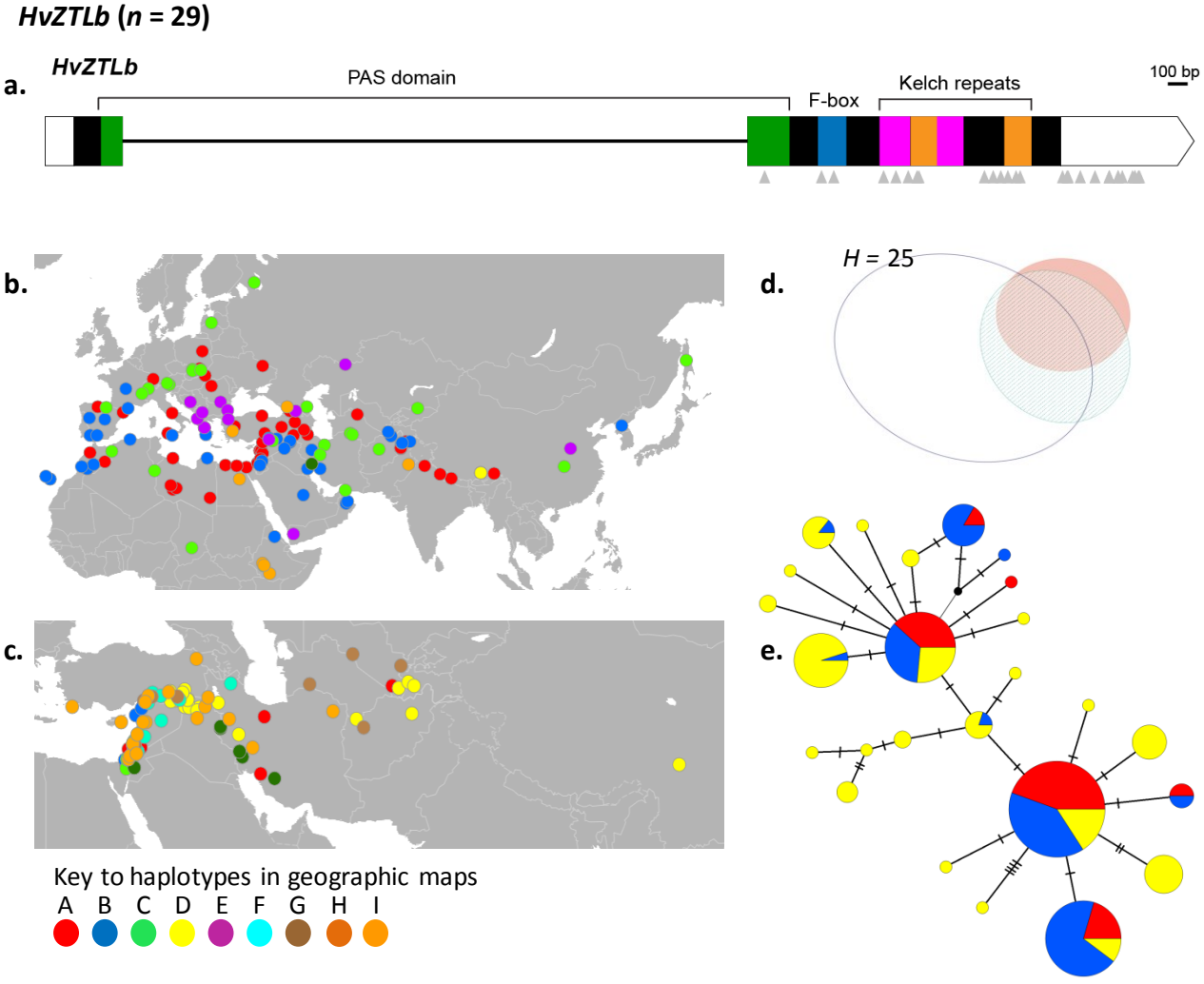
Supplementary Fig. 27: a. Genomic structure of *HvPRR95* and positions of the 69 SNPs identified. Synonymous SNPs and SNPs in the 3'UTR or introns are indicated by small grey arrows. SNP 43, which removes PTC of intron 6 (making it in frame on I6R alternative transcripts), and non-synonymous SNPs are represented by large black arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (Pseudo-Response Regulator) or green (CCT domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



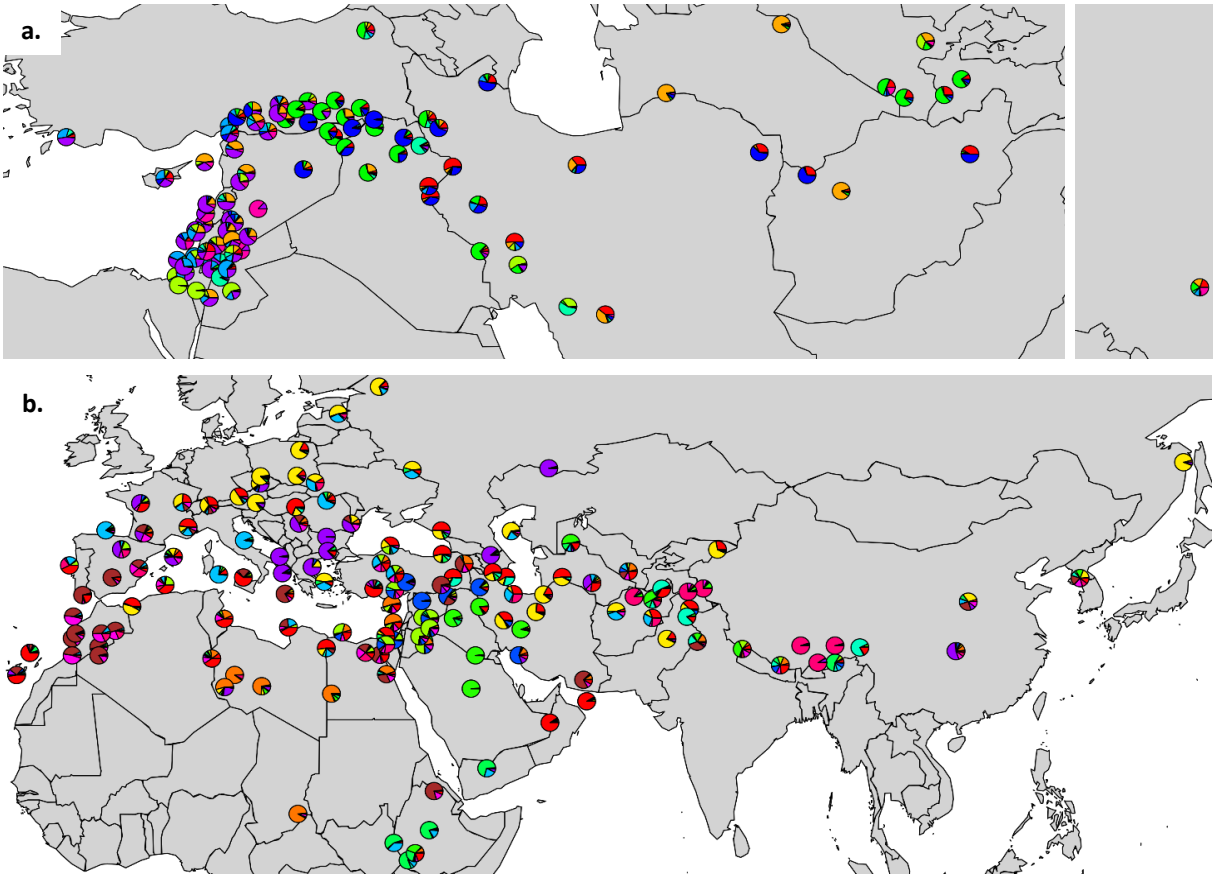
Supplementary Fig. 28: **a.** Genomic structure of *HvTOC1* and positions of the 83 SNPs identified. Synonymous SNPs and SNPs in UTRs are indicated by small grey arrows. Non-synonymous SNPs are represented by large colour coded arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded red (Pseudo-Response Regulator) or green (CCT domain) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



Supplementary Fig. 29: **a.** Genomic structure of *HvZTLa* and positions of the 49 SNPs identified. Synonymous SNPs and SNPs in the 3'UTR are indicated by small grey arrows. Non-synonymous SNPs are represented by large colour coded arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded green (PAS domain), blue (F-box-type domain), or orange/pink (Kelch motifs) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



Supplementary Fig. 30. **a.** Genomic structure of *HvZTLb* and positions of the 29 SNPs identified. Synonymous SNPs and SNPs in the 3'UTR are indicated by small grey arrows. 5' and 3' UTRs are open boxes; coding sequences are dark boxes, except domain-encoding exons, which are shaded green (PAS domain), blue (F-box-type domain), or orange/pink (Kelch motifs) **b.** and **c.** haplotype distributions according to geography **d.** Venn diagrams of haplotype number and sharing between different groups **e.** Median joining network. Further details given in general legend.



Supplementary Figure 31. Fractional sNMF assignments for 228 geo-referenced accessions based on all genic SNPs pooled across 19 flowering-associated genes. **a.** Spontaneum group accessions (*N* = 91, *K* = 9), **b.** Landrace group accessions (*N* = 137, *K* = 14) groups. The large interval between a single Chinese accession and other accessions (see also **Fig. 3**) has been truncated in the spontaneum map. Clear geographic structuring in both barley groups is evident, as observed for all SNPs. To visualise the majority of individual points, some positions are marginally offset.

Supplementary Figures references

1. Calixto, C.P.G., Waugh, R. & Brown, J.W.S. Evolutionary relationships among barley and *Arabidopsis* core circadian clock and clock-associated genes. *J. Mol. Evol.* **80**, 108-119 (2015).
2. Liu, T., Carlsson, J., Takeuchi, T., Newton, L. & Farré, E.M. Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. *Plant J.* **76**, 101-114 (2013).
3. Nakamichi, N. *et al*. *Arabidopsis* clock-associated pseudo-response regulators PRR9, PRR7 and PRR5 coordinately and positively regulate flowering time through the canonical CONSTANS-dependent photoperiodic pathway. *Plant & Cell Physiol.* **48**, 822-832 (2007).
4. Yoshida, R. *et al*. Possible role of EARLY FLOWERING 3 (ELF3) in clock-dependent floral regulation by SHORT VEGETATIVE PHASE (SVP) in *Arabidopsis thaliana*. *New Phytol*. **182**, 838-850 (2009).
5. Song, Y.H., Smith, R.W., To, B.J., Millar, A.J. & Imaizumi, T. FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. *Science* **336**, 1045-1049 (2012).
6. Sawa, M., Nusinow, D.A., Kay, S.A. & Imaizumi, T. FKF1 and GIGANTEA complex formation is required for day-length measurement in *Arabidopsis*. *Science* **318**, 261-265 (2007).
7. Sawa, M. & Kay, S.A. GIGANTEA directly activates *Flowering Locus T* in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **108**, 11698-11703 (2011).
8. Zhang, C. *et al*. Crosstalk between the circadian clock and innate immunity in *Arabidopsis*. *PLoS Path.* **9**, e1003370 (2013).
9. Staiger, D., Zecca, L., Wieczorek, K.D.A., Apel, K. & Eckstein, L. The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA. *Plant J*. **33**, 361-371 (2003).
10. Streitner, C. *et al*. The small glycine-rich RNA binding protein *At*GRP7 promotes floral transition in *Arabidopsis thaliana*. *Plant J*. **56**, 239-250 (2008).
11. Huang, N.-C., Jane, W.-N., Chen, J. & Yu, T.-S. *Arabidopsis thaliana CENTRORADIALIS homologue* (*ATC*) acts systemically to inhibit floral initiation in Arabidopsis. *Plant J.* **72**, 175-184 (2012).
12. Adams, S., Manfield, I., Stockley, P. & Carré, I.A. Revised morning loops of the *Arabidopsis* circadian clock based on analyses of direct regulatory interactions. *PLoS One* **10**, e0143943 (2015).
13. Turner, A., Beales, J., Faure, S., Dunford, R.P. & Laurie, D.A. The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Science* **310**, 1031-1034 (2005).
14. Jones, H. *et al.* Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent. *Mol. Biol. Evol.* **25**, 2211–2219 (2008).